Segmentation d'image par intégration itérative de connaissances

Mahaman Sani Chaibou Salaou

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THESE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, Image et Vision*

Par

Mahaman Sani CHAIBOU SALAOU

Segmentation d'image par intégration itérative de connaissances

Thèse présentée et soutenue à ISITCom Hammam-Sousse, le 02 Juillet 2019
Unité de recherche : ITI
Thèse N° : 2019IMTA0140

Rapporteurs avant soutenance :

Jean-Paul HATON	Professeur, Université Henri Poincaré, Nancy 1
Imed Riadh FARAH	Professeur, Université de la Manouba, Tunis

Composition du Jury :

Président :	Ouajdi KORBAA	Professeur, ISITCOM Hammam-Sousse
Examinateurs :	Jean-Paul HATON	Professeur, Université Henri Poincaré, Nancy 1
	Imed Riadh FARAH	Professeur, Université de la Manouba, Tunis
	Ali KHENCHAF	Professeur, ENSTA Bretagne
Dir. de thèse :	Basel SOLAIMAN	Professeur, IMT Atlantique Brest
Co-dir. de thèse :	Mohamed Ali MAHJOUB	Maître de conférences, ENISo Sousse
Encadrant:	Karim KALTI	Maitre-Asistant, FSM Monastir

Résumé

Le traitement d'images est un axe de recherche très actif depuis des années. L'interprétation des images constitue une de ses branches les plus importantes de par ses applications socio-économiques et scientifiques. Cependant cette interprétation, comme la plupart des processus de traitements d'images, nécessite une phase de segmentation pour délimiter les régions à analyser. En fait l'interprétation est un traitement qui permet de donner un sens aux régions détectées par la phase de segmentation. Ainsi, la phase d'interprétation ne pourra analyser que les régions détectées lors de la segmentation.

Bien que l'objectif de l'interprétation automatique soit d'avoir le même résultat qu'une interprétation humaine, la logique des techniques classiques de ce domaine ne marie pas celle de l'interprétation humaine. La majorité des approches classiques d'interprétation d'images séparent la phase de segmentation et celle de l'interprétation. Les images sont d'abord segmentées puis les régions détectées sont interprétées. En plus, au niveau de la segmentation les techniques classiques parcourent les images de manière séquentielle, dans l'ordre de stockage des pixels. Ce parcours ne reflète pas nécessairement le parcours de l'expert humain lors de son exploration de l'image. En effet ce dernier commence le plus souvent par balayer l'image à la recherche d'éventuelles zones d'intérêts. Dans le cas échéant, il analyse les zones potentielles sous trois niveaux de vue pour essayer de reconnaitre de quel objet s'agit-il. Premièrement, il analyse la zone en se basant sur ses caractéristiques physiques. Ensuite il considère les zones avoisinantes de celle-ci et enfin il zoome sur toute l'image afin d'avoir une vue complète tout en considérant les informations locales à la zone et celles de ses voisines.

Pendant son exploration, l'expert, en plus des informations directement obtenues sur les caractéristiques physiques de l'image, fait appel à plusieurs sources d'informations qu'il fusionne pour interpréter l'image. Ces sources peuvent inclure les connaissent acquises grâce à son expérience professionnelle, les contraintes existantes entre les objets de ce type d'images, etc.

L'idée de l'approche présentée ici est que simuler l'activité visuelle de l'expert permettrait une meilleure compatibilité entre les résultats de l'interprétation et ceux de l'expert. Ainsi nous retenons de cette analyse trois aspects importants du processus d'interprétation d'image que nous allons modéliser dans l'approche proposée dans ce travail :

1. Le processus de segmentation n'est pas nécessairement séquentiel comme la plus part des techniques de segmentations qu'on rencontre, mais plutôt une suite de décisions pouvant remettre en cause leurs prédécesseurs. L'essentiel étant à la fin d'avoir la meilleure classification des régions. L'interprétation ne doit pas être limitée par la segmentation.
2. Le processus de caractérisation d'une zone d'intérêt n'est pas strictement monotone i.e. que l'expert peut aller d'une vue centrée sur la zone à vue plus large incluant ses voisines pour ensuite retourner vers la vue contenant uniquement la zone et vice-versa.
3. Lors de la décision plusieurs sources d'informations sont sollicitées et fusionnées pour une meilleure certitude.

La modélisation proposée de ces trois niveaux met particulièrement l'accent sur les connaissances utilisées et le raisonnement qui mène à la segmentation des images.

Abstract

Image processing has been a very active area of research for years. The interpretation of images is one of its most important branches because of its socio-economic and scientific applications. However, the interpretation, like most image processing processes, requires a segmentation phase to delimit the regions to be analyzed. In fact, interpretation is a process that gives meaning to the regions detected by the segmentation phase. Thus, the interpretation phase can only analyze the regions detected during the segmentation.

Although the ultimate objective of automatic interpretation is to produce the same result as a human, the logic of classical techniques in this field does not marry that of human interpretation. Most conventional approaches to this task separate the segmentation phase from the interpretation phase. The images are first segmented and then the detected regions are interpreted. In addition, conventional techniques of segmentation scan images sequentially, in the order of pixels appearence. This way does not necessarily reflect the way of the expert during the image exploration. Indeed, a human usually starts by scanning the image for possible region of interest. When he finds a potential area, he analyzes it under three view points trying to recognize what object it is. First, he analyzes the area based on its physical characteristics. Then he considers the region's surrounding areas and finally he zooms in on the whole image in order to have a wider view while considering the information local to the region and those of its neighbors.

In addition to information directly gathered from the physical characteristics of the image, the expert uses several sources of information that he merges to interpret the image. These sources include knowledge acquired through professional experience, existing constraints between objects from the images, and so on.

The idea of the proposed approach, in this manuscript, is that simulating the visual activity of the expert would allow a better compatibility between the results of the interpretation and those of the expert. We retain from the analysis of the expert's behavior three important aspects of the image interpretation process that we will model in this work :

1. Unlike what most of the segmentation techniques suggest, the segmentation process is not necessarily sequential, but rather a series of decisions that each one may question the results of its predecessors. The main objective is to produce the best possible regions classification.
2. The process of characterizing an area of interest is not a one-way process i.e. the expert can go from a local view restricted to the region of interest to a wider view of the area, including its neighbors and vice versa.
3. Several information sources are gathered and merged for a better certainty, during the decision of region characterisation.

The proposed model of these three levels places particular emphasis on the knowledge used and the reasoning behind image segmentation.

Mots-clés

Image interprétation, Image segmentation, Superpixels, Mesure de similarité, Croissance de régions, Descripteurs contextuels, Descripteurs multi-niveaux, Apprentissage artificiel, Forêts aléatoires

Remerciements

Ce manuscrit présente mon travail de thèse effectuée au sein du laboratoire LATIS à l'ENISo et le département ITI à Brest. Ce fut une expérience, des plus riches, durant laquelle j'ai eu à côtoyer plusieurs personnes qui méritent pleinement ma gratitude.

Je tiens tout d'abord à exprimer ma haute reconnaissance à mon directeur de thèse Mohamed Ali MAHJOUB pour sa confiance dans l'orientation de mes travaux, son soutien et sa disponibilité constante pendant toute la durée de ce travail. À mon co-directeur, Basel SOLAIMAN, j'adresse mes sincères remerciements pour ses précieux conseils et ses encouragements tout au long de ce séjour scientifique. Merci également pour m'avoir toujours prêté une oreille attentive.

Je souhaite, ensuite, remercier mon encadrant Karim KALTI pour sa patience clairvoyante et son soutien quotidien inconditionné. Ses encouragements, ses discussions éducatives et sa grande disponibilité ont été pour moi un vrai tremplin lors de ces longues années turbulentes. Mes remerciements particuliers vont à Pierre-Henri CONZE, mon co-encadrant, pour ses explications adroites et ses remarques fines qui m'ont beaucoup édifié dans mes travaux de recherche. Ses relectures pertinentes ont constitué un apport considérable dans l'accomplissement de ce modeste travail.

Mes remerciements sadressent également à Ouajdi KORBAA pour l'honneur quil m'accorde en présidant mon jury de thèse.

J'exprime mes sincères remerciements aux professeurs Imed Riadh FARAH et Jean-Paul HATON pour avoir accepté d'être les rapporteurs de ce travail. J'aimerais ensuite exprimer toute ma gratitude au professeur Ali KHENCHAF pour avoir accepté d'assister à mon jury en tant qu'examinateur de ce travail.

Je remercie très vivement Mme Najoua ESSOUKRI BEN AMARA d'avoir accepté mon invitation pour assister à la présentation de ce travail. Par son biais, je remercie tous les membres du laboratoire LATIS pour leur convivialité et leur présence.

Enfin, je remercie tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce travail.

Dédicaces

À la mémoire de ceux qui nous ont précédé,
et à ceux qui ont cru à l'impossible.

Sommaire

Liste des figures

Liste des tableaux

Introduction générale

Ce chapitre présente d'abord la problématique de ce travail de thèse en exposant son cadre applicatif, les problèmes qui y sont abordés ainsi que ses principaux objectifs. Il présente également, de manière sommaire, la façon dont les problèmes sont abordés suivie des solutions proposées.

Scientists generally agree that no theory is 100 percent correct. Thus, ***the real test of knowledge is not truth, but utility****.*

Yuval N.H.

Sommaire

1. Cadre général

Dans le domaine de la vision par ordinateur, l'interprétation est définie comme le processus par lequel une scène, représentée par une ou plusieurs sources d'information, est décrite par son contenu sémantique de manière automatique par un système informatique. Ce processus a fait l'objet de beaucoup de modélisations théoriques comme en témoigne la pléthore d'approches proposées dans la littérature et implémentées sous forme de systèmes informatiques. Ces systèmes, destinés à des experts, sont utilisés sous forme d'outils d'aide à la décision dans divers domaines tels que la médecine, le militaire ou l'agriculture. Dans l'interprétation des mammographies, par exemple, les systèmes d'analyse détectent automatiquement les zones *suspectes* dans les images pour permettre aux radiologues de se focaliser directement sur les zones détectées. En effet, certaines de ces zones sont souvent omises par le radiologue car elles sont de petites tailles ou difficilement identifiables. La conduite assistée est un autre cas d'utilisation des systèmes d'analyse automatique d'images. Au moyen des caméras embarquées sur l'automobile, le système analyse la scène captée par les caméras et fournit des suggestions de directions au conducteur. Ces applications sont très populaires dans l'industrie automobile.

Dans la plupart des approches, le processus d'interprétation fait intervenir plusieurs outils issus de la vision par ordinateur, dont les plus fondamentaux concernent la segmentation et la reconnaissance d'objets. En effet, ces deux opérations analysent et extraient les objets de la scène pour générer une description interprétant la scène étudiée. En particulier, la segmentation permet le passage d'une image de pixels à une image de régions, dont la sémantique et la granularité dépendent de la méthode utilisée. Cependant, pour un expert humain, la segmentation n'est pas une finalité, mais plutôt une étape intermédiaire qui permet de passer à l'interprétation. Les deux étapes sont faites simultanément, d'ailleurs la segmentation peut être guidée par la sémantique que l'interprétation attribue aux zones -partiellement- reconnues. L'interprétation est alors un processus incrémental qui itère alternativement entre **focalisation** d'attention et **propagation** des connaissances jusqu'à stabilité. La focalisation d'attention permet l'identification des zones reconnues de l'image tandis que la propagation permet de progressivement répandre les connaissances courantes sur les reconnues zones vers le reste de l'image (zones incertaines). La figure .1 présente une description schématique du processus générique d'interprétation de scènes.

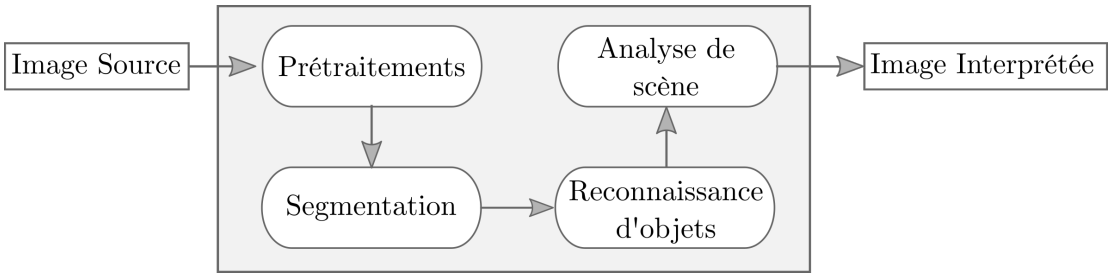


FIGURE .1. – Chaine classique du processus d'interprétation d'image.

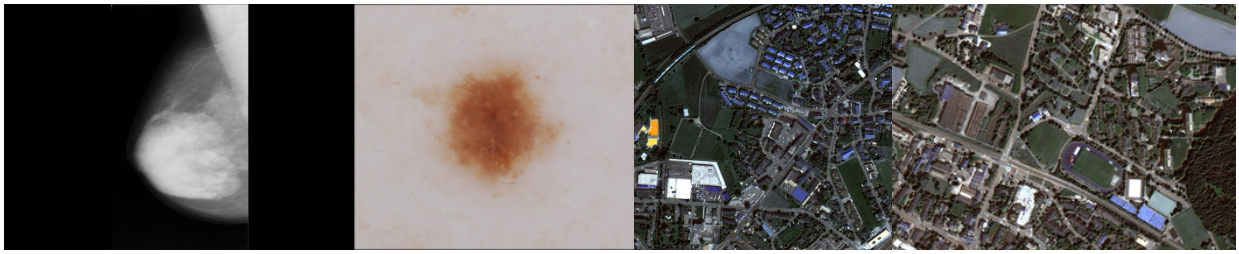
Habituellement, les systèmes d'interprétation intègrent deux principaux types de connaissances : Les connaissances sur le domaine d'application de l'image à interpréter tel que la taille ou la position relation des objets retrouvés dans les images et les connaissances relatives au domaine du traitement d'images

(indépendamment du domaine d'application considéré) tel que les propriétés des zones de contours ou des formes prédéfinies. Les connaissances du domaine sont nécessaires pour l'interprétation, mais aussi peuvent servir à guider l'utilisation des connaissances issues des traitements.

D'un point de vue ingénierie logiciel et systèmes à base de connaissances, deux catégories de systèmes d'interprétation se distinguent.

- Les systèmes dédiés. Ils sont conçus pour traiter un type d'images bien particulier. Dans ce type de systèmes les connaissances du domaine d'application et celles du traitement d'images sont généralement mélangées ce qui rend leur adaptation à d'autres domaines d'application difficile.
- Les systèmes non-dédiés. Ce sont des systèmes qui peuvent facilement s'adapter à plusieurs domaines d'applications. En effet, ils offrent une intégration séparée des deux types de connaissances afin de permettre un changement du domaine d'application à moindre coût. Cet aspect générique de ces systèmes pose, de manière accrue, de nombreux défis quant à la mise en uvre de ces systèmes, du fait de la complexité croissante des domaines d'application abordés. Les principales difficultés sont ainsi dues à l'intégration des connaissances relatives au domaine d'images à traiter plutôt que celles du traitement d'images lui-même. Dans ce contexte, la littérature des systèmes d'interprétation présente diverses formalisations de ces connaissances. Chaque modèle selon les données traitées conduit à une ou plusieurs approches d'intégration des connaissances dans le processus d'interprétation.

Cette thèse s'inscrit dans ce cadre général d'analyse automatique d'images. Plus particulièrement, nos travaux s'intéressent à l'intégration des connaissances au sein des méthodes de segmentation à des fins d'interprétation, dans des domaines où le contenu des scènes analysées ne varie pas trop tels que le domaine médical et celui de la télé-détection. Il s'agit de proposer une approche de segmentation non-dédiée d'images de scènes fixes ou à faible variation. En considérant l'aspect incrémental du processus d'interprétation, l'approche à proposer doit être facilement intégrable dans une approche d'interprétation afin de tirer profit des connaissances générées lors des étapes intermédiaires de la phase d'interprétation. Quelques exemples des images des scènes visées par ce travail sont illustrées dans la figure .2.



(a) (b) (c) (d)

FIGURE .2. – Quatre images des scènes visées par ce travail. Les deux premières illustrent des exemples de scènes médicales, mammographies dans .2a et mélanome dans .2b. Les deux dernières montrent des scènes de télédétection.

2. Objectifs et défis

La plupart des travaux d'intégration des connaissances dans les approches d'interprétation de scènes n'utilisent les connaissances que pour sélectionner l'algorithme de segmentation à appliquer afin d'extraire une primitive visuelle. Ces connaissances ne sont pas utilisées dans le processus de segmentation en soi.

L'interprétation nécessite la segmentation pour aboutir. Néanmoins les résultats de la segmentation peuvent être améliorés par ceux de l'interprétation, même intermédiaires. Dans notre travail l'objectif premier est la définition d'une **approche de segmentation** permettant l'**intégration transparente des connaissances** au sein du processus d'interprétation de scènes fixes ou à faibles variations. Ce modèle de segmentation doit garantir une intégration précoce des connaissances dans le processus. La spécification d'un tel modèle appelle, dans un premier lieu, à une formalisation abstraite des connaissances ainsi qu'à une définition d'un mécanisme de raisonnement sur ces connaissances, dans l'optique de *piloter* les étapes de la segmentation. La formalisation abstraite des connaissances et le raisonnement apportent une généricité ponctuelle au modèle. Plus particulièrement ce travail vise une représentation des connaissances sur plusieurs niveaux de sémantique correspondants aux différents points de communication entre l'étape de segmentation et celle d'interprétation. Cela fournira un transfert harmonieux des connaissances entre segmentation et interprétation. L'intégration des connaissances, au fur et à mesure de leur découverte, dans la segmentation est un processus complexe dont l'approche doit maintenir l'homogénéité des résultats tout au long de la tâche. En effet, le processus de segmentation doit être constamment réajusté de façon automatique afin de garder la cohérence entre les différentes itérations de la segmentation. Le tableau .1 résume les principaux objectifs visés par ce travail ainsi que les défis à relever.

L'intégration des connaissances entre l'étape de la segmentation et celle de l'interprétation nécessite une formalisation des connaissances qui soit compréhensible par les deux étapes. Bien que l'approche se focalise sur des scènes assez précises, la variabilité et les imperfections des connaissances peuvent rendre cette tâche complexe. À cette complexité viennent s'ajouter les difficultés inhérentes à la tâche de segmentation telles que l'évolution sémantique des entités manipulées (traduisant le passage des pixels vers des régions voire des objets de la scène). Sachant que les connaissances sont utilisées pour permettre à la segmentation de produire des résultats plus précis, les opérations élémentaires de la segmentation seront alors définies sur des données abstraites mais en gardant un comportement cohérent indépendamment des entrées reçues.

TABLE .1. – Tableau des principaux objectifs et défis du travail présenté.

Objectifs	**Défis**
Modèle de représentation de connaissances indépendant de l'application	Manipulation abstraite de connaissances. Pilotage de la segmentation par des connaissances abstraites.
Modèle de raisonnement indépendant de l'application	
Intégration précoce des connaissances dans la segmentation	Pilotage du comportement de la segmentation (e.g. raffinement des résultats).

Lorsque la scène est décrite par le seul contenu de l'image, l'interprétation de scène est équivalente à celle d'image. Dans la suite de ce manuscrit, nous parlerons de ces deux expressions de façon interchangeable car ce travail s'intéresse à des scènes décrites par une seule image.

3. Approche de segmentation proposée

Bien que l'objectif de l'interprétation automatique soit d'avoir le même résultat qu'une interprétation humaine, la logique des techniques classiques de ce domaine ne marie pas celle de l'interprétation humaine. La majorité des approches classiques d'interprétation d'images séparent la phase de segmentation et celle de l'interprétation. L'approche présentée dans ce travail tente de simuler l'activité visuelle de l'expert lors du processus d'interprétation pour permettre une meilleure compatibilité entre les résultats de l'interprétation et ceux de l'expert. Ainsi, ce travail propose une approche de segmentation d'images pour une intégration dans un processus d'interprétation utilisant une approche itérative et incrémentale. Nous focalisons notre étude sur des scènes fixes ou à faible variation telles que l'interprétation en radiologie ou en télé-détection. Cette approche s'intéresse plus précisément aux connaissances utilisées lors de cette analyse et à la manière de leur intégration dans le processus de la segmentation. Il s'agit tout d'abord de la représentation des connaissances puis du raisonnement sur ces connaissances pour aboutir à une segmentation qui pourra être complétée ensuite par une interprétation de l'image.
En général, les systèmes de vision ou d'analyse sémantique distinguent communément quatre niveaux d'abstraction de connaissances permettant d'aboutir à l'interprétation d'image. Dans l'ordre ascendant de sémantique, ces niveaux sont définis comme suit :

- Le niveau **Pixel** : il représente le niveau le plus élémentaire et le plus bas de la hiérarchie. Le pixel contient l'information extraite du capteur physique de l'image.
- Le niveau **Région** : une région est le résultat du regroupement d'un ensemble de pixels selon des critères de ressemblance purement physique.
- Le niveau **Objet** : ce niveau décrit les objets pouvant être présents dans l'image. Ils sont formés par regroupement des régions du niveau précédent.
- Le niveau **Scène** : c'est le niveau le plus haut. Il permet une description sémantique du contenu de l'image. Il s'agit par exemple de détecter des comportements, des situations, des types de zones ou encore des évènements.

Dans le cadre de ce travail, nous ajoutons un cinquième niveau, entre le niveau **Région** et celui **Objet**, que nous désignons par niveau **Classe thématique**. Sachant qu'une région n'est qu'un groupement de pixels voisins, le niveau **Classe thématique** regroupe alors les régions qui sont étiquetées avec un -type d'- objet de la scène. Ce sont en général des parties homogènes d'objets de la scène. D'autre part, nous proposons une approche de segmentation par intégration itérative des connaissances basée sur les régions. Notre approche implémente deux modèles d'intégration de connaissances comme suit :

- Formalisation explicite des règles de mise à jour des connaissances : ce modèle est appliqué au niveau de la croissance de régions.
- Construction des règles de mise à jour par apprentissage machine : ce modèle est utilisé pour la formation des objets de la scène.

Cette approche se limite aux quatre premiers niveaux (pixel, région, classe thématique, objet) de représentation des connaissances et se compose de trois grandes étapes (décomposition en superpixels, croissance de régions, formation des objets) qui traduisent le passage entre les niveaux d'abstraction considérés. La figure .3 présente un aperçu sommaire de l'approche de segmentation proposée montrant les niveaux de connaissances considérés.

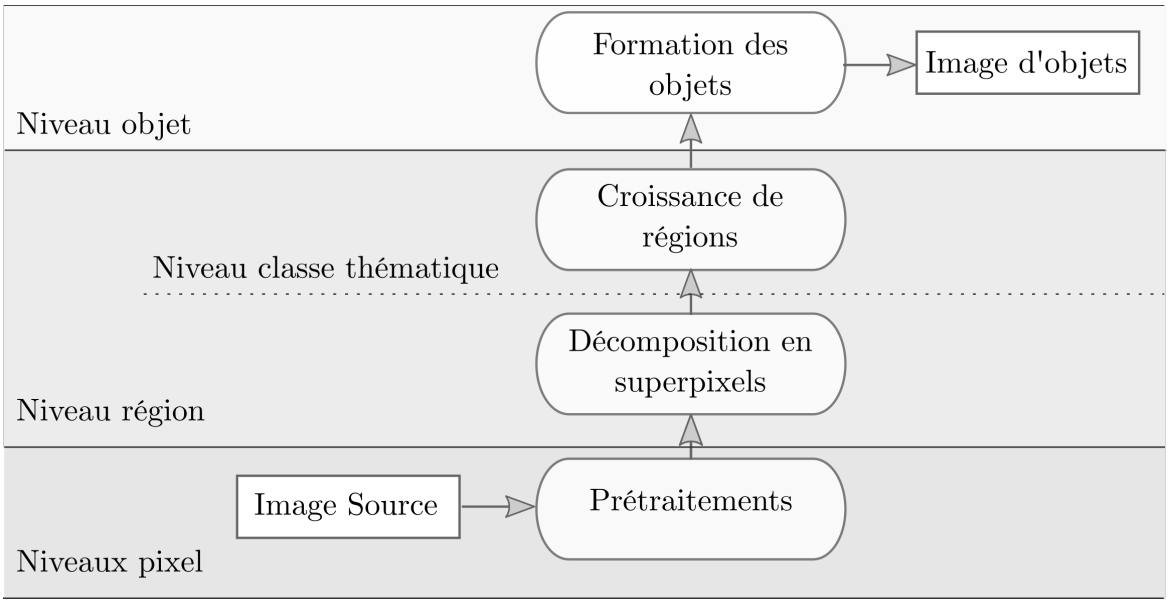


FIGURE .3. – Aperçu sommaire de l'approche de segmentation proposée exposant les niveaux de connaissances considérés.

3.1. Passage du niveau Pixel vers le niveau Région

Les pixels de l'image seront regroupés à l'aide de la méthode de segmentation en superpixels (Ren et Malik, 2003), en un ensemble de `N` régions de petites tailles et de niveau de gris similaires (presque identique). Cette méthode, appelée décomposition en superpixels, a la particularité de produire des régions compactes atomiques, de par leurs propriétés physiques. L'avantage du passage par cette étape est le gain qu'on réalise en traitement sans perte du côté de l'information. Il est à noter qu'il est possible d'ignorer cette étape et commencer directement à partir du niveau pixel. Dans ce cas, chaque pixel sera considéré comme une région élémentaire de départ.

3.2. Passage du niveau Région vers le niveau Classe thématique

Les superpixels issus de l'étape précédente représentent les premières régions obtenues par l'approche. Chaque région est caractérisée par des descripteurs physiques parmi lesquels la moyenne et la dispersion de ses niveaux de gris. Ainsi, soit $C = \{C_1, C_2, \ldots, C_K\}$ l'ensemble de classes sémantiques pour le type d'images considéré et préalablement définies. Les régions, obtenues suite à la décomposition en superpixels, sont ensuite étiquetées selon les classes C_k par classification selon leur taux d'appartenance à chacune de ces classes. Ainsi, après classification, chaque région R_i sera représentée par un vecteur donnant ses taux d'appartenances aux différentes classes thématiques C_k. $\tau_i = (\tau_{i1}, \tau_{i2}, \ldots, \tau_{iK})^T$ où τ_{ik} représente le taux d'appartenance de la région R_i à la classe C_k.

Le résultat de cette étape est donc une matrice de partition $\mathcal{R} = (R_1, R_2, \ldots, R_N)$ dans laquelle chaque région R_i est représentée par son vecteur de taux d'appartenance τ_i. La segmentation et l'interprétation sont réalisés par l'intermédiaire de cette matrice au lieu des valeurs numériques des pixels de l'image. Cette transition nous permet de travailler dans un espace sémantiquement plus riche et ainsi de réduire le fossé sémantique dans les phases suivantes du processus. La segmentation opère par croissance de région et procède d'une manière itérative en trois phases : la **focalisation**, l'**intégration et la propagation** des

connaissances et enfin la **fusion** des régions.

3.2.1. Focalisation

L'étape de la classification permet de caractériser les régions par leur information sémantique selon les différentes classes précédemment définies. Tandis que certaines régions se retrouvent divisées entre plusieurs classes, d'autres ont la particularité d'appartenir presque exclusivement à une seule classe. La focalisation permet de déterminer les zones de l'image par lesquelles le traitement sera initié : il s'agit de choisir les régions les mieux classifiées de l'image, aussi appelées régions germes. Cela nous assure de commencer le processus avec la meilleure certitude possible.

3.2.2. Intégration et propagation des connaissances

Après focalisation sur les zones reconnues de l'image, cette étape d'intégration et propagation des connaissances consiste à utiliser l'information sémantique des régions germes pour aider la caractérisation des régions voisines moins certaines. Il est question de *propager* la connaissance des régions germes vers leurs voisines afin d'améliorer leur classification sémantique. Cette étape constitue l'une des phases importantes de l'approche proposée. En effet, elle définit le procédé de mise à jour des taux d'appartenance des régions en prenant en considération pour chacune d'elles les trois sources de connaissances suivantes : le contexte local, le contexte global et le contexte sémantique de l'image.

3.2.3. Fusion

Dans une approche de segmentation de type croissance itérative de régions, la fusion est une étape nécessaire qui réalise le regroupement des régions afin de former les régions de l'itération suivante. En effet, la mise à jour des propriétés sémantiques des régions peut conduire à l'augmentation de la similarité de certaines régions connexes. Ainsi, cette étape de regroupement assure la fusion des régions selon des critères de similarité et la re-caractérisation des nouvelles régions obtenues. Les critères de similarité sont définis en fonction du niveau de représentation de connaissances courant.

Le cycle Focalisation - Intégration et propagation - Fusion est itéré jusqu'à la stabilité des régions dans l'image.

3.3. Passage du niveau Classe thématique vers le niveau Objet

Les régions du niveau **Classe thématique** se caractérisent principalement par une homogénéité visuelle. Elles sont produite suite à des fusions successives selon des critères purement numériques. Le niveau objet définit l'ensemble des objets de la scène observée. Cet ensemble peut se diviser en deux catégories :

- les objets **simples** : ils sont formés d'une seule région visuellement homogène et sont détectés à l'issue des fusions du niveau Classe thématique.
- les objets **complexes** : ils sont constitués par plusieurs régions hétérogènes. Leur formation nécessite une étape supplémentaire qui permet de regrouper les régions hétérogènes appartenant au même objet sémantique. Cette sémantique est intégrée au moyen des connaissances *à priori* sur les objets à détecter.

4. Organisation du manuscrit

La suite de manuscrit est organisée en cinq chapitres comme suit :

- Le chapitre « **Segmentation pour l'interprétation de scène** » présente un état-de-l'art sur les approches et les systèmes d'interprétation de scènes. L'objet du chapitre étant d'étudier les approches de segmentation dans l'interprétation de scène, une attention particulière est portée d'abord sur les *connaissances* et leur intégration dans ces approches. Ensuite, l'*utilité* ainsi que les *techniques d'utilisation* de ces connaissances dans la phase de segmentation sont détaillées. Le chapitre se termine par un inventaire critique non exhaustif des systèmes d'interprétation de scène mettant particulièrement l'accent sur les connaissances utilisées et leur intégration.
- Le chapitre « **Segmentation par propagation des connaissances** » présente une vue globale sur l'approche de segmentation proposée. Plus spécifiquement, ce chapitre traite de la représentation des connaissances et du raisonnement sur ces connaissances pour aboutir à une segmentation qui pourra être suivie par une interprétation de l'image. Ainsi, l'approche de segmentation proposée se compose d'une étape de **Focalisation** d'attention suivie de celle de **Propagation et intégration** de connaissances, et enfin d'une étape de **Fusion** des régions. Ces trois étapes sont successivement répétées jusqu'à stabilisation des régions.
- Le chapitre « **Croissance des régions adaptative** » aborde l'aspect itératif de cette approche. Ainsi, il présente une méthode adaptative de croissance des régions permettant de mieux contrôler l'évolution des régions pour offrir des résultats plus précis de la segmentation. D'autre part, un nouveau critère de similarité inter-régions est introduit. Cette contribution renforce le contrôle sur la formation des régions finales de la segmentation.
- Le chapitre IV, intitulé « **Similarité des superpixels par apprentissage** », présente une nouvelle méthode d'intégration de connaissances utilisant le regroupement sémantique de régions afin de former les objets complexes. Dans une image, ce type d'objets est généralement composé de plusieurs parties visuellement hétérogènes (voiture, bâtiment,. . .). Sachant que les approches de segmentation classiques ne peuvent pas atteindre un tel niveau de partitionnement de l'image, nous proposons dans ce chapitre d'utiliser une technique d'apprentissage machine pour permettre la formation d'objets complexes.
- Enfin, les conclusions sur l'ensemble des travaux ainsi que des perspectives qui pourront les compléter sont données dans le chapitre « **Conclusions** ». Ce chapitre récapitule et discute en outre les principales contributions issues des travaux de thèse.

La figure .4 situe graphiquement les principales contributions présentées dans les différents chapitres de ce manuscrit par rapport au modèle global.

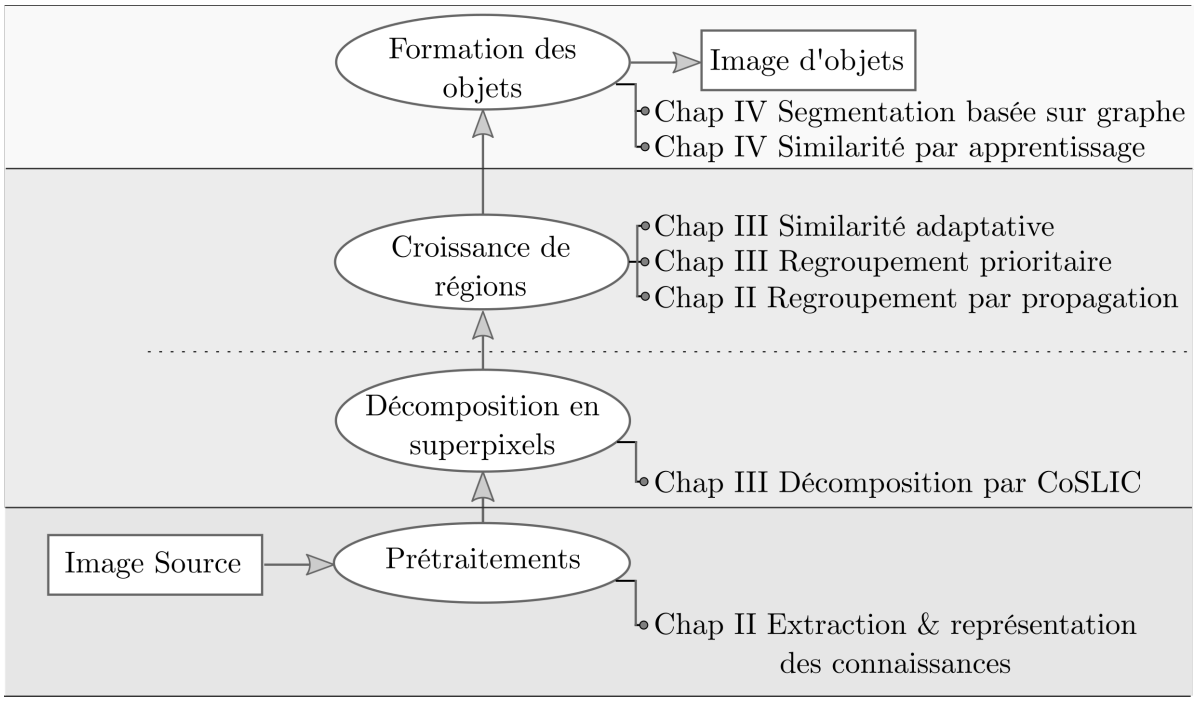


FIGURE .4. – Fil conducteur des principales contributions apportées dans ce travail.

I. Segmentation pour l'interprétation de scène

Dans le chapitre précédent, nous avions introduit de manière générale le travail exposé dans ce manuscrit. Ce chapitre présente un état-de-l'art des approches et des systèmes d'interprétation de scène. Tout d'abord, il énonce les approches d'interprétation en insistant sur les connaissances et leur intégration dans ces approches. Ensuite, l'utilité ainsi que les techniques d'utilisation de ces connaissances dans la phase de segmentation sont détaillées. Puis, il expose un inventaire critique non exhaustif des systèmes d'interprétation de scène mettant particulièrement l'accent sur les connaissances utilisées et leur intégration. La dernière partie du chapitre présente une analyse des méthodes de segmentation intégrées au processus d'interprétation.

Sommaire

1. Introduction

Le système humain de vision pour analyser une image se base sur les zones qui lui sont familières, et de là deviner les zones restantes en propageant les connaissances des zones reconnues. Un système d'interprétation de scène mets en œuvre deux éléments clés :

i) les **connaissances** qu'il intègre et

ii) les **traitements** qui utilisent ces connaissances dans l'atteinte des objectifs du système.

Les connaissances couvrent toute *information* extérieure fournie au système et sont exploitées afin de guider les traitements. La couleur d'un objet ou sa taille par rapport à un autre en sont des exemples. Le système peut aussi extraire des connaissances à partir des images traitées ou les générer au cours des traitements. Plusieurs critères sont utilisés dans la littérature pour caractériser les connaissances utilisées dans les systèmes d'interprétation de scènes : le **type**, le **paradigme**, les **niveaux de représentation** et l'**architecture de contrôle** des connaissances.

L'ensemble des traitements qui manipulent ces connaissances constituent le *raisonnement* du système. Il définit le modèle suivant lequel le système utilise ses connaissances. Dans le contexte du diagnostic du cancer du sein, l'utilisation de la taille pour différencier entre une masse et une calcification est un cas particulier qu'on retrouve dans les systèmes utilisés en mammographie. Il est défini par la stratégie de contrôle et l'architecture du système.

Ce chapitre présente un état-de-l'art des approches et des systèmes d'interprétation des scènes avec une attention particulière portée sur les *connaissances* et leur intégration dans ces approches. Après une définition et un positionnement des connaissances dans les approches d'interprétation, leur utilité ainsi que les techniques permettant leur utilisation dans la phase de segmentation sont détaillées (Sect.2.1). Ensuite, nous dressons un inventaire critique non exhaustif des systèmes d'interprétation de scène mettant particulièrement l'accent sur les connaissances utilisées ainsi que leur intégration (Sect.2.3). Le chapitre se termine par une analyse des méthodes de segmentation de type croissance de régions (Sect.3), qui sont intégrables dans le processus d'interprétation et une section de conclusion récapitulative (Sect.4).

2. Interprétation de scènes

Tout système d'analyse automatique de données qui se veut efficace doit intégrer des connaissances relatives aux données qu'il traite. Cette intégration peut prendre plusieurs formes selon les propriétés des connaissances impliquées. Elle définit l'approche du raisonnement du système d'interprétation.

2.1. Connaissances en interprétation de scènes

La connaissance peut se définir comme toute *action, fait de comprendre, de connaître les propriétés, les caractéristiques, les traits spécifiques de quelque chose*. Cette définition est reprise comme étant un ensemble de *faits et d'heuristiques* en vision par ordinateur dans (Harmon et King, 1985). Indépendamment de la définition, les travaux de la littérature évoquent plusieurs types de connaissances pour caractériser leur utilité dans le système.

2.1.1. Types de connaissances

De manière générique, dans le contexte d'un système de traitement d'images, les connaissances se divisent en deux principales catégories : les connaissances portant sur les opérateurs de traitement d'images et les connaissances relatives au domaine d'application considéré. La première catégorie de connaissances représente toute connaissance se rapportant au type d'images sur lequel s'applique le traitement. La seconde catégorie couvre toutes les connaissances permettant de caractériser les opérateurs de traitement d'images. Une troisième catégorie peut être ajoutée à cette liste : les connaissances stratégiques. Elles interviennent dans la stratégie d'agencement des opérateurs pour composer le traitement souhaité.

Cette première catégorisation des connaissances est axée sur le mode de leur utilisation dans le système sans tenir compte des caractéristiques intrinsèques à celles-ci. Ainsi, leur catégorisation selon des critères quantifiables est tout aussi importante afin de juger de l'importance des connaissances à intégrer dans un futur système. Cette étude se focalise essentiellement sur les connaissances du domaine d'application. Dans ce sens, plusieurs catégorisations des connaissances, résumées par la figure I.1, ont été réalisées selon les objectifs du système.

- Une des premières classifications des connaissances dans le domaine de l'interprétation d'images organise les connaissances en deux grandes catégories : les connaissances **implicites** et les connaissances **explicites** (Nikolopoulos *et al.*, 2011). Cette classification permet de caractériser l'extensibilité du système. Dans un système d'interprétation de scènes, les connaissances dites implicites sont directement codées par la logique du programme. Ces connaissances sont exprimées à travers la manière de résolution du problème posé. Cette intégration, très rigide, de connaissances ne prévoit aucune mise à jour des connaissances considérées. Aussi, ces connaissances sont fortement liées au domaine traité car changer de domaine nécessite de re-implémenter le système en entier.
  À l'inverse, une intégration explicite des connaissances permet une représentation des connaissances dans des structures autorisant une séparation entre les connaissances et les traitements qui les utilisent. Cette structuration des connaissances, souvent en plusieurs niveaux (§ Sect.2.1.3), offre divers avantages tels que l'extensibilité des connaissances du systèmes par la simplicité d'ajout de nouvelles connaissances sans changer les traitements associés.
- Selon leur nature, les connaissances peuvent aussi être divisées en connaissances **déclaratives** et connaissances **procédurales**. Les connaissances de type déclaratif permettent de définir les objets du monde réel par leur propriétés intrinsèques alors que les connaissances procédurales spécifient les liens entre ces objets ou leurs différents états. Ainsi, dans le système d'interprétation de scènes de rues (Matsuyama et Hwang, 1985), les auteurs dressent une séparation nette entre les caractéristiques des objets (fenêtres, portes, bâtiments,. . .), par les slots, et les relations régissant le regroupement de ces objets dans le contexte considéré, représenté par les prédicats.
- Mais, la répartition la plus courante des connaissances en interprétation de scènes distingue les trois classes suivantes : **temporelles**, **spectrales** et **spatiales**. Les connaissances temporelles définissent les évolutions possibles, dans le temps, des objets cibles. Les connaissances spectrales servent à décrire les objets considérés par leurs propriétés colorimétriques telles que la couleur ou la texture de ces objets. Enfin, les connaissances spatiales, parfois appelées aussi topologiques, ou contextuelles offrent une description des configurations spatiales des objets possibles dans le type d'images considéré.

Cette notion de topologie a été étendue par un quatrième type de connaissances, utilisé par Garnesson *et al.* (1992) et McKeown *et al.* (1985) pour caractériser un objet par sa fonction dans le monde de l'interprétation. Ainsi, ils regroupent plusieurs objets spatialement voisins et ayant une fonction commune en ce qu'ils appellent une *aire fonctionnelle*.

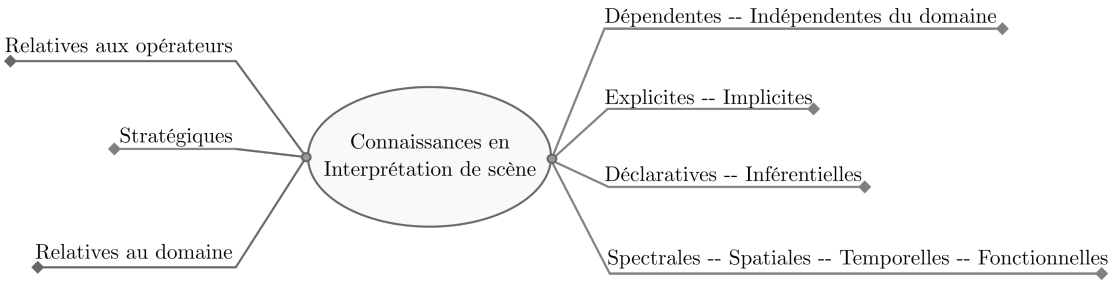


FIGURE I.1. – Types de connaissances : la catégorisation des connaissances rencontrées dans les systèmes d'interprétation de scènes.

Il faut noter que, dans tous les cas, on retrouve une séparation entre les connaissances relatives à la scène qui sont considérées comme de haut niveau et les connaissances relatives à l'objet, ou aux (groupes de) pixels dans l'image. La section 2.1.3 présente plus de détails sur cette organisation. Les connaissances permettent de décrire de manière sémantique les primitives extraites des images. Cette description est réalisée au moyen des formalismes de représentation afin d'être utilisable dans le système. Indépendamment de leur classe dans un système d'interprétation de scènes, les connaissances sont décrites par les trois propriétés suivantes : le formalisme, le niveau de représentation et la stratégie de contrôle.

2.1.2. Formalisme de représentation

En interprétation d'images, une bonne représentation de connaissance doit décrire à la fois les objets du monde réel ainsi que les actions possibles sur ces objets Rao et Jain (1988). Les paradigmes de représentation définissent les structures permettant de décrire les connaissances dans un système. Sept grandes catégories de formalismes de représentation de connaissances sont évoquées dans la littérature : la logique formelle, les règles de production, les frames, les réseaux sémantiques, les représentations orientées objet et les représentations basées sur les agents et les ontologies (Crevier et Lepage, 1997). La plupart des systèmes utilise une combinaison de plusieurs formalismes.

- La **logique formelle** est un mécanisme simple mais puissant de représentation de connaissances fournissant divers avantages mathématiques. Cependant, l'aspect monotone de la base de connaissances fait qu'il n'y ait aucun moyen de focalisation sur une connaissance ; toute recherche doit impérativement balayer toute la base (Crevier et Lepage, 1997 ; Harmon et King, 1985 ; Rao et Jain, 1988 ; Reiter et Mackworth, 1989).
- Les **systèmes de production** permettent de décrire les objets et leurs relations par la paire préconditions-actions (Levine *et al.*, 1978 ; Liedtke *et al.*, 1997 ; Matsuyama et Hwang, 1985 ; Newell, 1973). Les pré-conditions définissent les contraintes nécessaires pour former un objet ou une relation entre objets tandis que les actions décrivent les traitements permettant de générer l'objet ou la relation cible. Cependant des contradictions entre les règles peuvent facilement apparaître à cause de la mise à jour fréquente des règles.

- Les **frames** sont proposés Minsky (1975) et permettent une représentation de situations stéréotypées ou des classes d'objets (Draper *et al.*, 1989 ; Ikeuchi et Kanade, 1988 ; Iwase *et al.*, 1988 ; Matsuyama et Hwang, 1985). Ils présentent les connaissances sous deux niveaux hiérarchiques : (1) les données ou les faits, qui indiquent l'information qui est toujours vraie et (2) les slots qui décrivent les caractéristiques à préciser pour chaque instance de la situation. Cependant la coordination temporelle, les traitements en boucles, l'ordre ou le branchement sont très difficiles à implémenter (Crevier et Lepage, 1997).
- Aussi appelés graphes relationnels valués, les **réseaux sémantiques** (Liedtke *et al.*, 1997 ; Niemann *et al.*, 1990 ; Winston, 1970) permettent une représentation des connaissances en utilisant des nuds pour représenter les objets et les arcs matérialisent les relations sémantiques entre deux objets. Un réseau sémantique est simple et se traduit facilement dans un langage de programmation, néanmoins sa complexité augmente très vite avec le nombre de nuds.
- L'**approche orientée objet** quant à elle permet de se focaliser sur la modularité et favorise le partage des connaissances sur leur classification Crevier et Lepage (1997). Aussi, cette façon de faire assure une structuration rigide des connaissances à travers des mécanismes très simples tels que l'instanciation ou l'héritage (Clarkson, 1992 ; Forte *et al.*, 1992 ; Ikeuchi et Kanade, 1988).
- L'**ontologie** est la théorie des objets en termes de critères qui permettent de distinguer différents types d'objets et leurs relations, leurs dépendances et leurs propriétés (Town, 2006). Les ontologies encodent la structure relationnelle des concepts que l'on peut utiliser pour décrire et raisonner sur des aspects du monde. Cela les rend appropriés à la représentation des connaissances dans beaucoup de tâches dans l'analyse d'images (Maillot *et al.*, 2004 ; Popescu *et al.*, 2007 ; Town, 2006 ; Wache *et al.*, 2001).
- Un agent est "*un simple processus, qui, combiné avec d'autres agents produit des phénomènes complexe*" Minsky (1975). En interprétation d'images, les **agents** peuvent à la fois modéliser les objets et les tâches de traitement d'images. Ainsi dans (Niemann *et al.*, 1990), les auteurs proposent des agents pour intégrer les tâches de trois algorithmes pour segmenter les images. On retrouve une autre forme d'agents dans les travaux de (Draper *et al.*, 1989) où les agents émettent des hypothèses sur la partie de l'image qu'ils traitent et construisent des stratégies pour vérifier leurs hypothèses. Ces agents communiquent à travers le mécanisme de tableau noir *blackboard*. Ce dernier offre un espace partagé de collaboration où les différents agents peuvent lire les données et écrire les résultats de leurs tâches.

L'interprétation d'images par des systèmes intelligents peut être vue comme une correspondance entre la description de la scène et la structure de l'image (Matsuyama et Hwang, 1990). Cette correspondance nécessite une description progressive des connaissances allant des propriétés physiques de l'image aux objets réels présents dans la scène. La section suivante (§ Sect. 2.1.3) présente une synthèse des différents niveaux de description de connaissances retrouvés dans la littérature.

2.1.3. Niveaux de représentation

L'objectif principal de l'intégration des connaissances dans l'analyse d'images est de réduire le fossé sémantique en mappant les objets du monde réel vers une description bas niveau sur la machine. Les formalismes de représentation des connaissances présentés précédemment illustrent les différentes manières

de capter ces connaissances afin de les utiliser dans les systèmes d'analyse d'images. Mais ces différents paradigmes n'indiquent pas le découpage permettant le passage du monde réel vers l'adaptation de ces connaissances sur la machine. Il faut alors définir un modèle structuré détaillant clairement les niveaux de représentation des connaissances indépendamment du formalisme de représentation utilisé. Différents modèles de niveaux de description ont été proposés au cours des années.

- Kanade (1977) dans son système propose cinq niveaux hiérarchiques, pour une approche basée sur les régions : le *pixel*, le *patch*, la *région*, l'*objet* et la *sous-image*. Dans cette configuration, les niveaux pixel et patch utilisent des connaissances de bas niveaux, relatives aux propriétés purement physiques des données, en sortie du capteur d'acquisition. Les niveaux sous-image et objets sont appelés hauts niveaux et font intervenir la sémantique du domaine d'application, alors que le niveau région est considéré comme intermédiaire. Le niveau sous-image est constitué de groupements d'objets dans l'image.
  Une autre organisation est proposée dans **S**ystem for **P**hoto interpretation of **A**irports using **M**APS (SPAM) McKeown *et al.* (1985) où les auteurs considèrent les niveaux *région*, *segment*, *aire fonctionnelle* et *modèle*.
- Les segments, dans SPAM, représentent une région interprétée et les aires fonctionnelles sont constituées de segments remplissant une fonction commune dans le domaine concerné. Le système produit en sortie un ou plusieurs modèles d'interprétation de l'image caractérisé(s) par une organisation spatiale valide des aires fonctionnelles détectées.
  Plus simplement, certains travaux définissent les trois principaux niveaux suivants : le *pixel*, l'*objet* et la *scène*.
- On retrouve souvent cette configuration dans les systèmes utilisant les réseaux sémantiques (Liedtke *et al.*, 1997 ; Niemann *et al.*, 1990) et les frames (Matsuyama et Hwang, 1985) ; cela se justifie par le fait que le formalisme utilisé pour la représentation des connaissances inclut des mécanismes (les procédures pour les réseaux sémantiques et les slots pour les frames) qui incorporent entièrement les niveaux pixels, région et patch, en toute transparence.

En résumé, les niveaux de description des connaissances fournissent une définition des *unités de connaissances* manipulables dans le système considéré. Ainsi, une catégorisation, en quatre couches principales, de ces niveaux peut être identifié comme suit : la couche *pixel*, la couche *région*, la couche *objet* et la couche *scène*.

2.1.4. Stratégie de contrôle

La stratégie de contrôle se définit comme la façon d'utiliser les connaissances pour efficacement construire les objectifs du système. Selon Kanade (1977), elle est fortement liée au type de connaissances exploitées . Trois familles de stratégie sont à distinguer : bottom-up, top-down et mixte.
Un système d'interprétation de scène effectue deux tâches principales : la segmentation et l'interprétation. La stratégie de contrôle consiste à définir l'ordre d'intégration de ces tâches. Dans la stratégie du Bottom-up la segmentation vient avant l'interprétation tandis que dans le top-down cet ordre est inversé.

- **Bottom-up (data-driven) :** il s'agit d'une analyse se basant sur les données de l'image pour aboutir à une interprétation. C'est une stratégie qui est adoptée par la plupart des systèmes d'interprétation de scène. En effet, les systèmes commencent par analyser les caractéristiques numériques du contenu

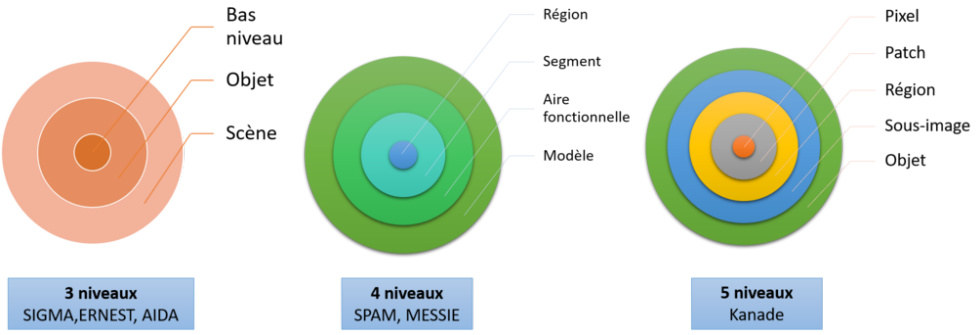


FIGURE I.2. – Trois types de niveaux de description de connaissances

des images pour ensuite faire correspondre des modèles sémantiques du monde réel. Dans leur système, Barrow et Tenenbaum (1976) segmentent l'image en associant des étiquettes aux régions obtenues. Une phase de vérification géométrique entre les étiquettes permet d'éliminer les étiquettes incohérentes. Mais cette stratégie n'est pas assez réaliste, ou du moins ne correspond pas au mode de fonctionnement du système visuel humain.

- **Top-down (goal-driven) :** dans cette stratégie, les systèmes partent de l'hypothèse qu'ils extraient des connaissances relatives à la scène pour prédire les caractéristiques physiques des zones concernées. Cette stratégie offre plus de garantie sur la cohérence des résultats de l'interprétation. Le raisonnement est fait des connaissances haut niveau vers les propriétés bas niveau en utilisant des conjectures appelées buts. Les systèmes dans cette catégorie commencent par définir des objectifs (buts) puis les vérifient ensuite en segmentant la zone visée comme c'est le cas dans (Bolles, 1976) et (Barrow et Tenenbaum, 1976).
- **Mixte :** de plus en plus de systèmes mélangent les deux stratégies selon les connaissances disponibles (Liedtke *et al.*, 1997 ; Matsuyama et Hwang, 1985). En effet, il est fréquent que les données ne soient pas suffisantes pour correspondre à la description d'un modèle sémantique d'une part. D'autre part, il est plus fiable dans certaines situations de caractériser les zones par rapport à leur entourage.

2.2. Approches d'interprétation de scènes

Une approche d'interprétation de scènes sous-entend la définition du problème de vision de la scène à interpréter et la manière avec laquelle ce problème est résolu. La suite de cette section présente une étude comparative des différentes approches phares d'interprétation proposées dans la littérature. Cette section commence avec des approches traditionnelles telles que celle de Marr (1982) pour finir avec d'autres visions un peu plus modernes comme la vision par objectif ou "purpose vision" [1].

2.2.1. Vision écologique

La théorie de la vision écologique est basée sur les travaux de Gibson (1979). Cette approche souligne la relation entre le système et son environnement. Elle souligne l'importance de l'environnement, de la nature de la lumière et du but des invariants dans la vision. La vision écologique suppose que les rayons

1. Dans le cadre de la vision par ordinateur, les expressions "perception visuelle" et "vision artificielle" peuvent être utilisées de manière interchangeable.

lumineux contiennent directement toutes les informations nécessaires à la reconnaissance du monde réel. Cette approche réfute l'utilisation des connaissances a priori et minimise l'importance du traitement de l'information et des représentations internes. Les tendances récentes sur la vision basée sur l'apparence sont basées sur cette théorie. Selon la vision écologique, le mouvement de l'observateur en impliquant un changement du flux optique permet de percevoir le monde. De plus la fonction d'objet a une grande importance sur la perception visuelle. Selon cette théorie, la sémantique associée à l'objet est relative à ses fonctions.

2.2.2. Paradigme de Marr

Au début des années quatre-vingts, Marr proposa une approche de vision par ordinateur couvrant le processus complet d'interprétation de scène. Ce paradigme approche le problème de la perception visuelle du point de vue de traitement de l'information. Il présente le processus de vision artificielle et son implémentation dans un système d'interprétation de scènes trois stades suivants :

- **la théorie computationnelle :** elle décrit ce que le système est supposé faire, les types de sorties selon les types d'entrées mais aussi quelles sont les traitements nécessaires associés.
- **les représentations et les algorithmes :** ce stade couvre l'aspect logiciel du processus. Il regroupe les structures de représentations des données utilisées ainsi que les algorithmes qui les manipulent.
- **l'implémentation :** elle représente le point de vue matériel du processus. Elle détermine les spécifications physiques des algorithmes.

Le paradigme de Marr présente la vision artificielle n'est qu'une succession bottom-up de processus qui transforme une information d'un niveau d'abstraction à un autre niveau d'abstraction plus haut. En projetant cette dernière définition dans le cadre du traitement de l'information image, la vision artificielle revient à une succession de la *segmentation*, de la *reconstruction* et de la *reconnaissance* des formes sur l'image considérée. C'est ainsi qu'il identifie les trois niveaux de représentation suivants de l'information qui traverse ce pipeline :

- **le croquis initial (primal sketch) :** ce niveau détermine les structures de l'images ainsi que les changements pouvant se produire. Il s'agit entre autre des variations de l'intensité du niveau de gris, des textures, des formes, etc.
- **le croquis 2.5D (2.5D sketch) :** c'est un niveau intermédiaire entre le croquis initial et le niveau 3D de la scène. Il couvre les géométries des surfaces visibles de l'image.
- **la représentation 3D :** c'est la vue tridimensionnelle de la scène indépendamment de l'observateur. Chaque objet est représenté de manière indépendante du point de vue.

Il faut noter que les deux premières représentations sont construites dans un repère centré par rapport au point de vue de l'observateur.
Le paradigme de Marr présente une vision hiérarchique intéressante de la vision artificielle qui non seulement simplifie la compréhension et l'implémentation mais aussi permet une séparation des différents niveaux de modélisation de la scène. Cependant, il faut noter quelques défauts inhérents à cette conception des choses. Plus particulièrement, il est impossible de reconstruire la scène exacte car les représentations sont, à la fin, faite de manière totalement indépendantes les unes des autres. L'absence d'intégration de connaissances *a priori* enlève toute possibilité d'interprétation "sémantique" de la scène. Aussi, l'interprétation serait

biaisée car la scène est vue sous l'angle de chaque objet de manière isolée sans proposer une vue globale de la scène considérée.

2.2.3. Vision active

Cette approche d'interprétation de scène, introduite par Aloimonos *et al.* (1988), présente la scène comme une zone dont l'observateur explore en ajustant sa vision pour mieux détecter certaines propriétés de la scène. Ainsi, selon ces auteurs, lorsqu'un être humain observe ses yeux ajuste la luminosité, le focus sur certaines choses afin d'obtenir une meilleure vue de la scène. Aloimonos *et al.* (1988) déclarent que les quatre problèmes fondamentaux suivants de la vision artificielle sont initialement mal posés dans les approches où les observateurs sont passifs.

- **la forme par l'ombrage (shape from shading) :** il s'agit de reconstruire la forme à partir de l'ombrage. Le problème consiste à déterminer, pour chaque point d'une image, la profondeur et/ou la surface normale du point correspondant sur la surface visible.
- **la forme par le contour (shape from contour) :** il est question d'étudier comment détecter la forme des objets à partir des contours reconnus dans l'image. Plusieurs sous problèmes peuvent en découler de cette définition tels que la *localisation* des contours, la *distinction* des contours de différents objets ou l'*interprétation* des contours détectés.
- **la forme par la texture (shape from texture) :** le problème de la reconstruction de la forme à partir de la texture consiste à retrouver une orientation de la texture permettant de recouvrir la surface de l'objet. Ainsi, à partir de cette surface sera reconstruite la forme dudit objet.
- **la forme par le mouvement (shape from motion) :** le problème est de récupérer le mouvement tridimensionnel et la structure d'un objet en mouvement à partir d'une séquence de ses images. Deux approches se distinguent : La première suppose un "petit" mouvement. La deuxième approche suppose que le mouvement est grand, et la mesure du mouvement de l'image entraîne la résolution du problème de correspondance.

Les auteurs de la vision active démontrent théoriquement que leur approche permet de mieux poser les quatre problèmes pré-cités et fournit, par conséquent, de meilleures appréhensions des problèmes menant à des solutions moins couteuses.
Cependant, il est plutôt clair que l'approche est plus dynamique que active. En effet, l'observateur bouge mais n'est pas actif. De plus, ces travaux sont restés trop théoriques.

2.2.4. Perception active

La notion de perception active a été introduite par Bajcsy (1988). La perception visuelle et l'interprétation sont définies comme un problème de contrôle. L'objectif est de planifier des stratégies de contrôle pour améliorer la connaissance du système sur son environnement et pour une acquisition intelligente des données. La perception active est définie comme l'étude de la modélisation des stratégies de contrôle de la perception visuelle. La modélisation affecte à la fois les capteurs et les modules de traitement. Cette modélisation se divise en deux, comme suit :

- **Modèles locaux :** ils représentent les paramètres des différents modules de traitement (paramètres du capteur, paramètres des algorithmes de traitement d'image). Ces paramètres permettent la prédiction du comportement et / ou des résultats des modules de traitement.

- **Modèles globaux :** ce sont les paramètres qui représentent l'interaction entre les différents modules, c'est-à-dire comment les différents modules sont fusionnés (superviseur). L'idée principale est l'introduction d'une boucle rétroactive dans le système. Cette boucle rétroactive permet au système d'acquérir des données uniquement lorsqu'elles sont nécessaires au système.

L'approche de la perception active est intéressante car elle prend en compte, de manière explicite, non seulement les représentations mais aussi les processus qui agissent sur ces représentations. La stratégie de perception consiste à rechercher la succession d'actions pour obtenir un maximum d'informations avec un coût minimal. Les travaux sur la perception active mettent l'accent sur (1) la représentation explicite à la fois des connaissances et du raisonnement, (2) la notion de processus de raisonnement et de contrôle, (3) l'importance de ne traiter les données que lorsqu'elles sont nécessaires.

2.2.5. Vision téléologique

La vision téléologique a été introduite par Aloimonos (1990). Dans cette approche, l'idée de base est de diviser le problème initial en sous-problèmes. Le travail consiste en la définition des modules de traitement dédiés à chaque sous-problème et de la définition d'un superviseur qui gère les différents modules. La division du problème global en sous-problèmes et leur regroupement en un module général permettent d'améliorer la perception visuelle et les tâches d'interprétation.
La vision téléologique a donné naissance à un large éventail de travaux et d'applications. Certaines notions fondamentales ont inspiré la recherche en vision cognitive. En particulier, cette approche insiste sur :

- le point de vue orienté vers la tâche de la perception visuelle et de l'interprétation,
- la notion de systèmes minimalistes : réaliser uniquement les tâches pertinentes pour atteindre objectif souhaité,
- la division de tâches complexes en sous-tâches plus faciles à traiter.

La plupart des systèmes de vision construits avec le paradigme de la vision téléologique dépendent cependant fortement de l'application.

2.2.6. Vision animée

La vision animée a été proposée par Ballard et Brown (1992). Cette approche est basée sur l'étude des mouvements intentionnels de l'il humain au cours d'une tâche visuelle. Ballard et Brown (1992) considèrent la perception visuelle et l'interprétation dans le contexte d'une action. Une représentation 3D du monde réel n'est pas nécessaire. De même que pour la vision active, cette approche considère la vision comme un problème mal posé. Le but de la vision animée est d'ajouter des contraintes grâce aux informations fournies par le mouvement contrôlé des capteurs. Le but de cette approche est de contrôler le mouvement du capteur pour atteindre le centre d'attention et regarder les tâches de contrôle. Cette méthode vise à réduire la complexité de la perception visuelle et des tâches d'interprétation. Un système de vision animé peut décaler les capteurs, changer la mise au point et l'angle de vision. La vision animée utilise un système de coordonnées exocentrique centré sur l'objet. Concernant la mise en uvre du système, Ballard et Brown (1992) utilise une tête binoculaire active.
L'idée principale de la vision animée est la notion de stratégie de recherche visuelle par la mise en place de mécanismes de contrôle du regard et de focalisation de l'attention dans l'image. Ces mécanismes permettent d'analyser uniquement les parties pertinentes des images.

2.3. Systèmes d'interprétation de scènes

L'interprétation de scènes par des systèmes intelligents s'articule autour de la représentation des connaissances de la scène à interpréter et de la méthode de raisonnement sur ces dernières. Cela confère à ces systèmes une séparation naturelle de ces deux principales notions. Pour la plupart, cette séparation est implémentée par des modules experts organisés selon le type de connaissances ou leur niveau de description. Le modèle de communication entre les principaux modules est appelé architecture ou organisation modulaire du système. On en distingue deux catégories : les systèmes hiérarchiques et les systèmes hétérarchiques.

Dans les systèmes hiérarchiques, les modules sont agencés selon une structure ascendante où les plus hauts *commandent* les plus bas. MESSIE (Garnesson *et al.*, 1992) et AIDA (Liedtke *et al.*, 1997) sont des exemples de système utilisant ce type d'architecture. Cependant dans une hétérarchie, les modules sont organisés selon une structure plate de manière à favoriser l'interrelation et la coopération comme c'est le cas dans les systèmes SPAM (McKeown *et al.*, 1985), SIGMA (Matsuyama et Hwang, 1985) et ERNEST (Niemann *et al.*, 1990).

Plusieurs systèmes d'interprétation d'image ont été proposés dans la littérature. Le type de connaissances utilisé fut le premier critère de distinction entre les systèmes. Ensuite des systèmes plus modulaires se basant sur des formalismes de représentation des connaissances indépendants du traitement apparaissent vers le début des années 70 avec les travaux de (Winston, 1970). Les sections suivantes présentent une synthèse de quelques systèmes existants. Plus particulièrement, pour chaque système la tâche de segmentation proposée est analysée ainsi que son intégration avec le reste du processus d'analyse.

2.3.1. SPAM

System for **P**hoto interpretation of **A**irports using **M**APS (McKeown *et al.*, 1985) est l'un des premiers systèmes "*complets*", proposé dans les années 80. Les auteurs proposent, à travers ce système, d'étudier l'usage des systèmes à base de règles dans le contrôle des traitements d'images et l'interprétation des résultats par rapport à un modèle de référence ainsi que la représentation de ce modèle dans une base d'images . Le système développé permet l'interprétation de scènes d'aéroports dont il distingue deux catégories : les aéroports civils ou commerciaux et les aéroports militaires. L'interprétation est réalisée par l'étiquetage des régions segmentées suivi d'un regroupement des objets identifiés en "*aires fonctionnelles*" elles-mêmes organisées en modèle global. SPAM utilise le mécanisme d'hypothèse-vérification qui consiste à émettre des hypothèses sur les propriétés des zones à partir de leur entourage ou des connaissances apriori. Ensuite ces hypothèses seront validées en calculant les propriétés effectives de la zone. Aussi, le système utilise une architecture en hétérarchie comprenant trois modules principaux : le module d'extraction des primitives physiques, le module de raisonnement à base de règles et le module de mise en correspondance entre les données images et le modèle symbolique. Une comparaison entre les résultats du système avec des images segmentées par un humain et des images segmentées automatiquement montre sa robustesse par rapport aux images entrées.

La conception modulaire de SPAM permet une intégration de la technique de segmentation indépendante de l'approche considérée. La segmentation d'images est définie comme le processus de génération de régions candidates pour le processus d'extraction de primitives. Ces primitives correspondent aux régions

qui ont été reconnues et interprétées par un programme, ou qui ont été définies par un humain, ou qui peuvent être caractérisées par un ensemble de position, forme, taille et propriétés spectrales. L'approche de segmentation automatique intégrée dans SPAM a été proposée dans (McKeown et Denlinger, 1984).

I l s'agit d'une approche par croissance de régions qui utilise une base de carte de correspondance image-plan pour guider la formation des régions. Le plan prédit l'apparence et la position approximatives d'une entité dans une image. Il prédit également la zone d'incertitude causée par des erreurs dans la correspondance image-plan. Le processus de segmentation recherche ensuite les régions d'image qui satisfont les critères d'intensité et de forme bidimensionnels. En commençant par une image des contours, le processus de segmentation forme les régions par fusions successives des régions séparées par le contour sélectionné. La sélection des contours est faite itérativement des plus faibles aux plus forts. Si aucune région initiale n'est trouvée, le processus tente de fusionner les régions pouvant satisfaire ces critères. Selon les types de régions recherchées plusieurs critères peuvent être générés. Par exemple, une combinaison des critères d'une certaine plage d'intensité moyenne, de surface, de compacité, de linéarité est utilisée lors de la recherche de zones de tarmac dans des scènes d'aéroports.

2.3.2. SIGMA

Proposé en 1985 par Matsuyama et Hwang, le système SIGMA (Matsuyama et Hwang, 1985, 1990) est un cadre d'interprétation de scènes aériennes caractérisé par une représentation des connaissances via les frames et les règles de production. En effet les connaissances sur les objets sont représentées par un frame dont les slots contiennent les propriétés physiques de l'objet ainsi que ses relations avec d'autres objets. En outre, des règles de production, sont aussi stockées dans les slots, composées de préconditions et des actions afin de définir les conditions nécessaires à un changement d'état de l'objet et l'effet de ce changement. Le système est divisé en trois principaux modules ou experts : l'expert de raisonnement géométrique (GRE), l'expert de sélection de modèle (MSE) et l'expert de vision bas niveau (LLVE). Pour interpréter l'image, les caractéristiques physiques sont, tout d'abord, extraites. Ensuite, une vérification des préconditions permet d'activer les objets "*évidents*" assez bien caractérisés. Une phase de reconnaissance est ensuite réalisée par un raisonnement spatial coopératif entre les différents experts. A noter que la notion d'objet partiellement reconnus permet à SIGMA d'implémenter les stratégies de Bottom-up pour caractériser les zones inconnues et Top-down pour chercher des éléments dans la zone de l'image qui entoure l'objet partiellement reconnu. Dans ce que les auteurs appellent le "*cycle d'interprétation*", ces objets sont ensuite vérifiés par souci de cohérence mutuelle et utilisés pour générer des hypothèses sur d'autres objets à extraire (Crevier et Lepage, 1997).

La segmentation d'images est réalisée par le module LLVE. Celui-ci utilise deux types de connaissances pour guider la segmentation :

- Connaissance des concepts fondamentaux de la segmentation d'images : types de primitives pouvant être extraites d'une image et types d'opérateurs de traitement d'images.
- Connaissance des techniques de segmentation d'images : comment combiner efficacement les opérateurs.

Le premier type de connaissance est représenté par un réseau décrivant le type de la structure dans la segmentation d'image et le dernier par un ensemble de règles de production. L'ensemble des types de

primitives que le système peut extraire est représenté par un réseau dans lequel les nuds sont les primitives et les arcs correspondent aux algorithmes qui permettent le passage entre primitives.

Pour une image $\mathcal{I}$, la segmentation est réalisée par découpage en plusieurs zones $\{\mathcal{I}_i\}$, généralement correspondant aux objets sémantiques. Ce découpage est produit par le module MSE. Chaque zone est analysée et segmentée à part à travers une séquence d'algorithmes experts d'extraction de primitives. Cette sequence est générée par le module LLVE à partir des objectifs de l'interprétation. Par exemple, afin d'extraire un bâtiment rectangulaire, une séquence de plusieurs d'extracteurs des lignes peut être générée qui seront ensuite regroupées en polygone puis en rectangle. Pour le même objectif, le module peut proposer un seul algorithme d'extraction de rectangle. Le module choisit, parmi l'ensemble des possibilités permettant d'atteindre l'objectif, la séquence au coût minimal pour générer la segmentation. Lorsque la séquence choisie n'aboutit pas, le module change les paramètres des algorithmes ou sélectionne une autre séquence, ainsi de suite jusqu'à obtention du résultat ou épuisement des séquences candidates. Le coût d'une séquence est calculé en fonction des coûts d'exécution des algorithmes et de leur "*utilité*". L'utilité permet de donner score d'efficacité à un algorithme dans un cas d'application bien précis. Par exemple l'extraction de contours n'a pas le même coût pour la même image en présence et en absence de bruits.

2.3.3. MESSIE

Dans leur système, dénommé **M**ulti **E**xpert **S**ystem for **S**cene **I**nterpretation and **E**valuation (MESSIE) (Garnesson *et al.*, 1989, 1992), Garnesson et ses co-auteurs abordent le problème de l'interprétation d'image à travers une architecture multi spécialistes (experts). Ils se focalisent sur l'expression des connaissances nécessaires à l'interprétation qu'ils considèrent comme la principale difficulté de l'interprétation de scènes. La représentation des connaissances adoptée est une association entre le paradigme orienté objet et les règles de production. Un objet sémantique est représenté par quatre vues : la forme géométrie, l'aspect (la radiométrie), le contexte spatial et la fonction (Sandakly et Giraudon, 1994). Aussi dans MESSIE une séparation des connaissances en deux niveaux est proposée : le niveau scène composé des connaissances sur la scène et des stratégies de contrôle et le niveau objet qui regroupe les connaissances relatives aux objets.

Le résultat de l'interprétation est une matrice de localisation fournissant un pointeur vers la position et les caractéristiques de chaque objet détecté. Les auteurs de MESSIE proposent, au niveau scène, une architecture hiérarchique de trois modules (le superviseur, le contrôleur de localisation et le contrôleur de scène) qui coopèrent à travers un tableau noir . Les modules du niveau objet sont des spécialistes d'extraction de caractéristiques physiques dédiés aux différents types d'objets supportés par le système. Les modules du système sont activés selon une stratégie Bottom-up lorsqu'un évènement jugé favorable à la caractérisation d'un objet sémantique est détecté dans le tableau noir. Aussi certains spécialistes, par exemple pour l'extraction des caractéristiques d'une région, sont parfois activés à la demande d'autres spécialiste plus "*intelligents*" à travers la stratégie Top-down.

MESSIE utilise deux types de segmentations : la segmentation contour et la segmentation région. La première génère des chaînes de contours (et les segments associés) et la seconde produit des régions qui sont exprimées sous forme de chaînes de contours ou de radiométrie et le graphe d'adjacence résultant.

- La segmentation contour est réalisée par composition séquentielle d'un filtrage gradient (Deriche, 1987), d'une suppression des non maxima locaux suivie d'un seuillage par hystérésis pour finir avec

un chainage des points de contours (Giraudon, 1987).

- Les régions sont produites par l'approche par croissance de régions proposée dans Giraudon et Montesinos (1987).

Les résultats des deux segmentations (chaines de contour et régions) sont des primitives qui sont stockés dans une structure de tableau noir (Hayes-Roth, 1983, 1985) sous forme d'objets structurés (avec des mécanismes d'héritage) pour permettre leur accessibilité aux autres systèmes experts du système global.

2.3.4. AIDA

En 1997, une équipe de l'université de Hannover en Allemagne propose un système d'interprétation de scène qu'ils appellent **A**utomatic **I**nterpretation of **DA**ta (AIDA) (Bückner *et al.*, 2001 ; Liedtke *et al.*, 1997, 2001). Ce système utilise des connaissances *a priori* sur les objets pour générer une interprétation de la scène observée sous forme de description symbolique des régions dans l'image. AIDA utilise une représentation explicite des connaissances sous forme de réseau sémantique contrôlé par des règles. Le système prend en entrée une image accompagnée de la connaissance apriori. A partir des connaissances, il génère une hypothèse sur le contenu qu'il vérifie par l'extraction des caractéristiques physiques de l'image ; il combine ainsi les deux stratégies de contrôle du top-down et du Bottom-up. Aussi notons qu'AIDA se divise en deux modules principaux : le module d'interprétation et le module de segmentation. Le système a été testé sur la reconnaissance de structures complexes, la recherche automatique des points d'ancrage pour le recalage des images de télédétection et la modélisation 3D des objets spécifiques du paysage et du bâtiment.

La segmentation se situe au bas niveau de représentation des connaissances. Comme dans les systèmes précédents, elle est réalisée à travers des opérateurs d'analyse d'images et sert principalement à extraire des primitives visuelles de la scène telles que les lignes ou les rectangles dans une zone sélectionnée de l'image d'entrée.

2.3.5. ROSESIM

Dans (Hudelot, 2005), les auteurs proposent un système d'interprétation sémantique d'images. L'objectif de ces travaux était de concevoir une plate-forme de vision cognitive réutilisable et générique pour le problème complexe de l'interprétation d'images sémantiques. Ils supposent une division du problème d'interprétation en trois sous-problèmes :

i) l'interprétation sémantique,

ii) la correspondance entre les représentations de haut niveau d'objets physiques et les données de capteurs extraites d'images,

iii) le problème de traitement d'image.

La plateforme proposée repose sur une architecture distribuée qui est basée sur la coopération de trois systèmes à base de connaissances (KBS). Chaque KBS est hautement spécialisé pour le sous-problème correspondant et dispose d'un moteur dédié et d'un modèle unifié de représentation des connaissances à travers des ontologies (Maillot *et al.*, 2004). Cette approche a été appliquée au diagnostic précoce des maladies des plantes, notamment les maladies des feuilles de roses dans les serres.

A l'instar des approche précédentes, celle de Hudelot (2005) ne fait pas une définition explicite de la tâche de segmentation. D'ailleurs, ces travaux portent plutôt un intérêt sur l'aspect génie logiciel et ingénierie de

connaissance (Hudelot et Thonnat, 2003). La détection et la formation des régions des images est ainsi divisée entre les deux modules inférieurs i.e. celui de gestion des données visuelles et celui du traitement d'images.

2.3.6. Fusion et interprétation possibiliste de scènes

Un système d'interprétation (PBL) de scène observée par de multiples capteurs utilisant la représentation et le traitement possibiliste des connaissances est présenté dans (Alsahwa, 2014). Cette approche permet d'effectuer une analyse hiérarchique de scène en se reposant sur deux processus : ascendant et descendant, illustré par la figure I.3. Le premier processus permet d'accumuler l'évidence sur l'existence des régions (ou objets), tandis que le second permet de remettre en cause, par l'expert, le contenu informationnel des régions et objets identifiés dans le processus ascendant. Visant des applications sur des scènes multi sources, les auteurs proposent d'exploiter les outils possibilistes pour permettre une fusion des différentes sources de connaissances et améliorer l'interprétation des données acquises en vue d'une représentation plus riche de la scène observée. Le système cible principalement la segmentation/classification (processus ascendant) et le démixage (processus descendant).

Notons que dans ses travaux Alsahwa (2014) met surtout l'accent sur la caractérisation et la fusion des connaissances qu'il réalise à travers le formalisme possibiliste. La tâche de segmentation résulte d'un processus itératif de diffusion croisée des connaissances possibiliste à partir d'une carte initiale générée par classification pixélique.

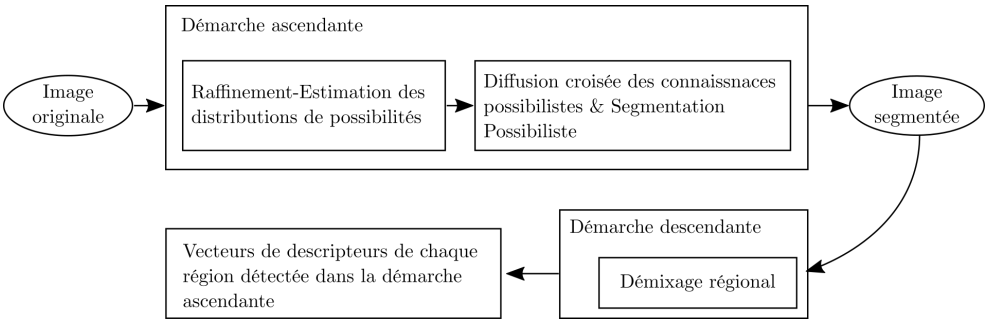


FIGURE I.3. – Architecture du modèle ascendant/descendant du système d'interprétation de scène basé sur une démarche possibiliste proposé par Alsahwa.

2.4. Synthèse analytique

Il est difficile de faire une situation précise de l'étape de segmentation dans le processus d'interprétation (Guigues, 2003) du fait même de sa définition imprécise. Le style de l'approche de segmentation utilisée dépend de l'approche d'interprétation et du système qui l'implémente. Cependant, relativement à l'échelle de description des connaissances (§ 2.1.3), l'étape de la segmentation est tant tôt définie dans le haut niveau (comme regroupement des éléments appartenant à un même objet sémantique) tant tôt dans le bas niveau (comme regroupement des éléments de même caractéristiques physiques). Néanmoins, nous pouvons dégager trois tendances à partir des approches exposées précédemment.

— La première vague regroupe les approches "très" théoriques qui laissent le choix de la technique de segmentation au système. Il s'agit de la Vision Écologique et de la Vision Active.

— La deuxième catégorie, sans spécifier une technique précise de segmentation, suppose l'obtention des résultats finaux issus de la segmentation afin de poursuivre avec d'autres tâches telles que la reconstruction dans le cas de la Vision de Marr ou la formation des objets dans le cas de la Vision Téléologique.

— Enfin la dernière tendance, observée dans la Perception Active et la Vision Animée, suggère l'utilisation d'une technique itérative pour alterner ses résultats intermédiaires avec ceux des autres tâches du processus. Cela leur permet de raffiner les résultats du processus au fur et à mesure des traitements.

	Caractéristiques Intrinsèques		**Segmentation**			
	Architecture	Application	Type	Intégration	Niveaux desc.	Contrôle
SPAM	Hétérarchie	Aéroports	Croissance de Régions	Indépendant	Région, Segment, Aire fonctionnelle, Modèle	Top-down
SIGMA	Hétérarchie	Images Aériennes	Multi Experts	Appel Procédure	Pixel, Patch, Région, Objet, Sous-image	Mixte
MESSIE	Hiérarchie	Images Aériennes	Multi Experts	Appel Procédure	Région-Segment-Aire fonctionnelle-Modèle	Top-down
AIDA	Hétérarchie	Images Aériennes	Multi Experts	Appel Procédure	Pixel-Région-Scène	Mixte
ROSESIM	Hiérarchie	Images Plantes	Multi Experts	Appel Procédure	Pixel-Région-Objet	Mixte
PBL	Hétérarchie	Images Médicales	Croissance de Régions	Indépendant	Sous-pixel-Pixel-Région-Objet	Bottom-Up

TABLE I.1. – Tableau comparatif des systèmes d'interprétation d'images.

Le tableau I.1 résume quelques propriétés distinctives des systèmes d'interprétation présentés. La grande majorité de ces systèmes traite des images de télédétection et présente une organisation modulaire hétérarchique. Dans tous ces systèmes la tâche de segmentation est reléguée à des composants élémentaires génériques et indépendants (entre eux, du système et des données d'entrée) appelés opérateurs. Le principal objectif est, étant donné une zone délimitée de l'image, d'extraire des formes régulières prédéterminées qui symbolisent les primitives visuelles de la scène qui sont supportées par le système hôte. La segmentation est donc essentiellement une tâche définie au bas de niveau qui ne traite que les données physiques de l'image. En général, ces algorithmes disposent d'une librairie d'opérateurs d'extraction de primitives visuelles -modules experts- qu'ils affectent à un objectif défini par le composant haut-niveau du système. Lorsqu'un opérateur échoue à extraire la primitive demandée, le système réessaie en changeant les paramètres de l'opérateur ou sélectionne un autre opérateur permettant d'atteindre le même objectif. Seuls les systèmes utilisant les approches de segmentation par croissance de régions (SPAM (McKeown *et al.*, 1985) et PBL (Alsahwa, 2014)) propose une modélisation explicite de la segmentation. Cependant, dans SPAM, la segmentation est définie au niveau région tandis que Alsahwa (2014) la définit au niveau *objet* sémantique, contrairement aux approches précédentes qui la restreignent à l'extraction de primitives élémentaires des images. Le travail de ce dernier constitue une partie de la base de l'approche proposée dans ce rapport.

La séparation explicite de la tâche de segmentation de celle d'interprétation, faite dans les systèmes d'interprétation, réduit considérablement l'importance de la segmentation. Les images sont d'abord

segmentées puis les régions détectées sont interprétées. Cependant, dans les approches d'interprétation, cette démarcation de la segmentation n'est nulle part évoquée. En effet, le raisonnement dans ces approches est fait en termes de régions sémantiques et non des lignes ou segments. Les primitives visuelles sont alors représentées à cheval entre le *bas* et le *haut* niveau de représentation. Ce nouveau positionnement de la segmentation montre que les approches de segmentation itératives et incrémentales, telles que celles du type croissance de régions, peuvent naturellement s'intégrer dans le processus d'interprétation. Les approches les plus récentes utilisent parfois l'expression de *segmentation guidée par les connaissances* lorsque des connaissances de haut niveau sont utilisées dans le processus.

3. Segmentation par croissance de régions

De manière générale, la segmentation est le processus qui permet de partitionner le contenu d'une image en des régions homogènes (Haralick et Shapiro, 1985). Pour un système d'analyse automatique d'images, la segmentation est une étape nécessaire et primordiale (Vincent *et al.*, 1985).

Les approches de segmentation par croissance de régions sont connues par leur principe simple et itératif. En effet, partant d'une image $\mathcal{I}$ à segmenter, il s'agit de sur-segmenter $\mathcal{I}$ en un ensemble de régions $\mathcal{R}$, ensuite par un processus itératif regrouper les régions de $\mathcal{R}$ selon des critères de similarité spatiale et colorimétrique (Schettini, 1993 ; Tremeau et Borel, 1997). Il faut noter que les premières approches proposées omettent l'étape de sur-segmentation en régions et par conséquent utilisent directement les pixels de l'image comme éléments à regrouper pour aboutir à la segmentation. Cette particularité leur offrent une bonne adhérence aux contours au détriment de l'efficacité et de la robustesse. En effet, avec les tailles des images toujours croissantes, l'efficacité est un critère qui détermine l'utilisabilité même des algorithmes de traitement d'images. Schématiquement, l'algorithme (Algo. I.1) des approches de segmentation par croissance de régions se résume comme suit :

1. **Initialisation :** ici, les pixels à partir desquels le processus de croissance sera initié sont choisis : ce sont les pixels germes.
2. Tant que non **convergence**, répéter :
   a) **Mesure de similarité :** cette tâche utilise les descripteurs des pixels pour calculer leur similarité visuelle avec les germes. Seuls les pixels spatialement connexes aux germes sont considérés.
   b) **Fusion de régions :** après le calcul des similarité, tous les pixels ayant une similarité supérieure à un seuil sont fusionnés pour croitre les régions germes.

Lorsqu'un algorithme dédié est utilisé pour choisir les régions à partir desquelles le processus itératif de regroupement est initié, on parle de croissance de régions avec sélection de (régions) germes (Adams et Bischof, 1994 ; Dreizin *et al.*, 2016 ; Huang *et al.*, 2018 ; Shih et Cheng, 2005 ; Thirumeni *et al.*, 2015). L'étape de sélection de germes permet d'orienter le processus de croissance de régions en priorisant le traitement des régions d'intérêt pour l'application considérée et repoussant celui du reste des régions.

3.1. Mesure de similarité et critère d'homogénéité

Dans une approche itérative de segmentation, la mesure de similarité entre entités est une étape cruciale. Elle assure le regroupement progressif des éléments afin d'obtenir les régions finales de la segmentation.

Algorithme I.1 : Segmentation par croissance de régions

```
Données : I, S_0, C_0 ;          ▷ I:Image source, S_0:Similarité de fusion, C_0:Condition d'arrêt
Résultat : R ;                                                  ▷ R:ensemble de régions
1 début
2     R := ChoixGermes(I) ;
3     répéter
4         pour R ∈ R faire
5             pour p ∈ N(R) faire
6                 si Similarité(R, p) ⩾ S_0 alors
7                     R := R ∪ p ;
8     jusqu'à Convergence(I, C_0) = Vrai ;
```

De manière générale, la similarité $S(X,Y)$ entre deux éléments X et Y exprime leur ressemblance et vérifie les trois axiomes suivantes :

— Non négativité : $S(X,Y) \geqslant 0$

— Symétrie : $S(X,Y) = S(Y,X)$

— Autosimilarité maximale : $S(X,X) = 1$

Dans le contexte de segmentation, en général les éléments correspondent aux régions à regrouper et sont représentées par leur caractéristiques extraites. Deux grandes familles de mesures de similarité ont été proposées, pour capter la similarité entre deux régions (Bouchon-Meunier *et al.*, 2008 ; Gan *et al.*, 2007) : les mesures similarités *géométriques* et les mesures similarités *ensemblistes*.

Les mesures similarités *géométriques* appelées aussi mesures de similarité métriques, considèrent les informations à comparer comme étant des points dans un espace métrique de représentation. Par conséquent, la similarité résulte de la comparaison des informations à travers une mesure de distance ensuite transforment cette distance en une mesure de similarité. Les distances métriques les plus connues utilisées pour la comparaison de données sont celles de *Minkowski* (De Amorim et Mirkin, 2012) et leurs dérivées (Hajdu et Tóth, 2008 ; Ichino et Yaguchi, 1994 ; Merigo et Casanovas, 2011). L'équation I.1 donne une formule générale d'une mesure de similarité basée sur ces distances, étant donné deux objets représentés par leurs vecteur de caractéristiques respectifs $X = \{x_i\}$ et $Y = \{y_i\}$, avec $i = 1, \ldots, \mathtt{N}$.

$$S_M(X,Y) = 1 - \frac{d_p(X,Y)}{\sqrt[p]{N}} \tag{I.1}$$

où $d_p(X,Y) = \sqrt[p]{\sum_{i=1}^{N} |x_i - y_i|^p}$ est la distance associée et $p \in \mathbb{N}$ correspond au facteur d'ordre de *Minkowski*. (Alsahwa, 2014) indique que les valeurs les plus utilisées du facteur d'ordre sont 1, 2, et ∞. Dans ces derniers cas les distances obtenues correspondent respectivement à la distance de Manhattan (Eq. I.2), celle Euclidienne (Eq. I.3) et de Cheybychev (Eq. I.4).

$$d_M(X,Y) = \sum_{i=1}^{N} |x_i - y_i| \tag{I.2}$$

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{N} |x_i - y_i|^2} \tag{I.3}$$

$$d_C(X,Y) = max_{i=1}^{N} |x_i - y_i| \tag{I.4}$$

En revanche, les mesures de similarité *ensemblistes* considèrent chaque information représentant un objet, comme étant un sous-ensemble algébrique de l'univers des primitives Ω, comportant les primitives réalisées par cette information. Ainsi, la similarité entre deux objets est calculée en fonction du nombre de primitives communes ou distinctes, aux informations représentant les objets à comparer. Dans cette catégorie, on retrouve des distances telles que l'*intersection normalisée* (Eq. I.6), l'indice de Jaccard (Jaccard, 1901) (Eq. I.7), le coefficient de *Sørensen − Dice* (Dice, 1945 ; Sørensen, 1948) (Eq. I.8). Plusieurs travaux proposent des extensions de ces mesures (Hadjieleftheriou et Srivastava, 2010 ; Paskaleva et Bochev, 2012). La distance proposée par (Tversky, 1977) est la plus généralement considérée pour une mesure de similarité de type ensembliste. Elle est formalisée comme suit :

$$S_{\alpha,\beta}(A,B) = \frac{|A \cap B|}{|A \cap B| + \alpha|A \setminus B| + \beta|B \setminus A|} \tag{I.5}$$

où $|\dot{|}$ calcule le cardinal de l'ensemble et $A \cap B$ représente le sous-ensemble des primitives communes à A et B. $\alpha \geqslant 0$ et $\beta \geqslant 0$ sont deux facteurs de pondération.

$$S_I(A,B) = \frac{|A \cap B|}{max(|A|,|B|)} \tag{I.6}$$

$$S_J(A,B) = \frac{|A \cap B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} \tag{I.7}$$

$$S_D(A,B) = \frac{|A \cap B|}{|A| + |B|} \tag{I.8}$$

A remarquer la relation $S_J = S_D/(2 - S_D)$ qui existe entre l'indice de $Jaccard$ et le coefficient de $Sørensen - Dice$ mais aussi $S_J = S_{\alpha=1,\beta=1}$ qui fait de l'indice de $Jaccard$ un cas particulier de $S_{\alpha,\beta}$. Les travaux de Jousselme et Maupin (2012) présentent une synthèse plus complète des distances de comparaison d'informations contenues dans deux ensembles.

Les mesures de similarité fournissent une valeur permettant de juger de la ressemblance de deux régions dans une approche de segmentation d'image par regroupement de régions. Dans beaucoup d'approche, un prédicat d'homogénéité $\hat{\mathcal{H}}$ locale est utilisé pour exprimer la "*fusionnabilité*" de deux régions adjacentes. Le critère d'homogénéité est généralement formé par la composition d'une ou plusieurs mesure de similarité sur les régions considérées.

Ainsi, Bins *et al.* (1996) proposent un critère d'homogénéité utilisant la distance Euclidienne calculée sur les valeurs des intensités des régions candidates. Le critère calculé est comparé à un seuil dynamique incrémenté à chaque nouvelle itération du processus de regroupement. Ce critère ne prend pas compte de la dimension spatiale des données image et produit des régions très irrégulières. Les travaux de (Baatz et Shäpe, 2000) introduisent cet aspect dans leur critère, connu sous le nom de critère de Baatz & Schäpe, en

ajoutant aux caractéristiques spectrales des attributs géométriques. Baatz et Shäpe expriment l'homogénéité de deux régions P et Q par :

$$f(P,Q) = \omega_c \times f_c + (1-\omega_c) \times f_s \tag{I.9}$$

f_c (I.10) et f_s (I.11) calculent respectivement l'hétérogénéité spectrale et spatiale entre P et Q. ω_c permet de contrôler l'importance relative des deux mesures intermédiaires f_c et f_s.

$$f_c(P,Q) = \sum_b \left[(\mathtt{A}_P + \mathtt{A}_Q) \times \mu_{P,Q_b} - (\mathtt{A}_P \times \mu_{P_b} + \mathtt{A}_Q \times \mu_{Q_b}) \right] \tag{I.10}$$

$$f_s(P,Q) = \omega_s \times d_c + (1-\omega_s) \times d_l \tag{I.11}$$

Tout comme la composante spectrale, celle spatiale utilise l'aire et le périmètre des régions pour calculer l'hétérogénéité de la régions résultante. Mais celle-ci ajoute une division par le rectangle englobant minimum pour ajouter un lissage spatial.

$$d_c(P,Q) = \mathtt{P}_{P,Q} \times \sqrt{\mathtt{A}_{P,Q}} - (\mathtt{P}_P \times \sqrt{\mathtt{A}_P} + \mathtt{P}_Q \times \sqrt{\mathtt{A}_Q}) \tag{I.12}$$

$$d_l(P,Q) = \frac{\mathtt{P}_{P,Q} \times \mathtt{A}_{P,Q}}{\mathtt{P}_{BB_{P,Q}}} - \left(\frac{\mathtt{P}_P \times \mathtt{A}_P}{\mathtt{P}_{BB_P}} + \frac{\mathtt{P}_Q \times \mathtt{A}_Q}{\mathtt{P}_{BB_Q}} \right) \tag{I.13}$$

où $\mathtt{A}_x$, $\mathtt{P}_x$, μ_{x_b} correspond respectivement à l'aire, le périmètre et l'écart-type des valeurs de l'intensité dans la bande spectrale b de région x. Au final, le critère de Baatz & Schäpe permet de prendre en compte l'aspect spatial en plus de celui spectral. Il nécessite trois paramètre et produit de bons résultats (Darwish *et al.*, 2003). Les auteurs de (Crisp *et al.*, 2003) proposent un autre critère qui utilise les attributs spectraux et spatiaux. Dans ce critère, appelé Full Lambda Schedule, l'homogénéité est définie par :

$$f(P,Q) = \frac{\frac{\mathtt{A}_P \times \mathtt{A}_Q}{\mathtt{A}_P + \mathtt{A}_Q} \times |\mu_P - \mu_Q|^2}{len(\partial(P,Q))} \tag{I.14}$$

$len(\partial(P,Q))$ calcule la longueur de la frontière commune entre P et Q. Une valeur de seuil trop grande produit des régions de tailles importantes mais partageant des courtes frontières. Dans le même sens les auteurs de (Peng *et al.*, 2011) introduisent un critère probabiliste qui calcule la cohérence entre deux régions via la formule de probabilité suivante :

$$\begin{cases} P_0(P,Q) = 1 - \lambda_1 exp(-(\mu_Q - \mu_P) \times S^{-1} \times (\mu_Q - \mu_P)) \\ P_1(P,Q) = 1 - \lambda_2 exp(-(\mu_Q - \mu_{P,Q}) \times S^{-1} \times (\mu_Q - \mu_{P,Q})) \end{cases} \tag{I.15}$$

où λ_1 et λ_2 sont des paramètres réels donnés. μ_x est la valeur moyenne des intensités de la région et S est la matrice de covariance des deux distributions. Ce critère utilisant un modèle de distribution Gaussien, estime le degré de cohérence de P et Q par P_0 et son inverse par P_1. Une étude plus approfondie des critères d'homogénéité entre ensemble est proposée dans (Calderero et Marques, 2010 ; Lassalle *et al.*, 2015).

3.2. Stratégie de fusion

Dans un processus de segmentation itératif par regroupement de régions, la mesure de similarité et le critère d'homogénéité permettent d'estimer la ressemblance visuelle de régions. Le choix des régions à

fusionner est assuré par la stratégie de fusion implémentée. Cette étape permet particulièrement, pour une région donnée, de sélectionner la meilleure région voisine parmi celles qui vérifient le prédicat du critère d'homogénéité. Différentes heuristiques pour la sélection d'une région adjacente ont été proposées ; (Baatz et Shäpe, 2000) en présente une synthèse intéressante.

Étant donné une région P et la liste de ses régions adjacentes, la stratégie de *Fitting (F)* consiste à choisir la première région adjacente Q pour laquelle le prédicat d'homogénéité $\hat{\mathcal{H}}$ est vérifié avec P. Une amélioration de cette première stratégie consiste à choisir une région adjacente Q pour laquelle le coût de fusion avec P est le minimum parmi ceux qui vérifient le prédicat. Cette heuristique est appelée *Best Fitting (BF)*. En considérant, BF et étant donné trois régions adjacentes, P, Q et R. Lorsque P choisit Q comme région de fusion et Q choisit R qui elle même choisit P, on se retrouve dans un cycle fermé infini. Ces régions ne vont jamais être regroupées dans ces conditions. L'heuristique *Local Mutual Best Fitting (LMBF)* ajoute la contrainte supplémentaire de sélection mutuelle. Cette contrainte stipule que deux régions P et Q sont regroupées si et seulement si elles vérifient le prédicat d'homogénéité et que P choisisse Q et Q choisisse P (Lassalle *et al.*, 2015).
Toutes ces stratégies définissent des conditions locales aux régions en cours de traitement. Cela permet de faire des regroupements en parallèle dans différentes zones de l'image. Cet avantage cependant soulève la question de gestion des conflits en cas de chevauchement de régions choisies. Pour pallier cette situation la dernière heuristique, *Global Mutual Best Fitting (GMBF)*, impose de ne fusionner qu'une seule paire de régions adjacentes à chaque itération du regroupement. La paire de régions qui fusionne est celle pour laquelle la mesure de similarité est la minimum parmi celles de toutes les autres paires de régions de l'image. Cette heuristique est la plus contraignante car seulement une paire fusionne à chaque itération, ce qui peut augmenter sévèrement le temps d'exécution de l'algorithme (Blaschke et Hay, 2001 ; Lassalle *et al.*, 2015).

4. Conclusion

L'interprétation de scènes est un processus automatique par lequel un système informatique associe à une image une description sémantique de son contenu. En partant de cette définition, nous avons fourni dans la première partie de ce chapitre une description générique des approches d'interprétation qui s'articule entre les connaissances et les raisonnements qui les manipulent pour produire l'interprétation des images. Ensuite, nous avons dressé un historique des approches et des systèmes bien connus de la littérature. La deuxième partie du chapitre nous a permis de discuter les approches de segmentation intégrées à celles de l'interprétation plus précisément la famille des approches par croissances de régions. En effet, nous avons détaillé les principales étapes constituant cette catégorie d'approches de segmentation tout en analysant les contributions faites dans la littérature pour chacune d'elles.

Le chapitre suivant de ce manuscrit expose notre proposition d'approche de segmentation de type croissance de régions. Nous présentons un nouveau modèle de croissance de régions qui procède par propagation itérative de connaissances.

II. Segmentation par propagation des connaissances

L'objet de ce travail de thèse est principalement la définition d'une approche de segmentation itérative qui permet l'intégration des connaissances spécifiques afin d'orienter progressivement le processus. Ce type d'approche s'intègre facilement dans un contexte d'interprétation de scène où elle pourra utiliser les connaissances disponibles et/ou générées par l'étape d'interprétation. Dans ce cadre, l'objectif est tout d'abord la représentation de ces connaissances puis le raisonnement sur ces connaissances pour aboutir à une segmentation qui sera suivie d'une interprétation de l'image. Ce chapitre présente une vue globale sur l'approche de segmentation proposée. Cette dernière est essentiellement composée d'une étape de ***Focalisation*** *d'attention suivie de celle de* ***Propagation et intégration*** *de connaissances, et enfin d'une étape de* ***Fusion*** *des régions. Ces trois étapes sont successivement répétées jusqu'à stabilisation des régions.*

Sommaire

1. Introduction

Le processus d'interprétation en vision par ordinateur permet, à partir d'une scène -qui est représentée par une image de ce qui est observé- de fournir une description du contenu de celle-ci. Comme nous l'avons présenté dans le chapitre de l'état-de-l'art (Chap. I), plusieurs approches ont été proposées au fur du temps. Chacune de ces approches est orientée par ses objectifs spécifiques et présente ses forces et ses faiblesses inhérentes. Mais le but commun de toutes ces approches était d'approcher la vision humain dans le cadre d'une interprétation de scène.

L'approche proposée dans ce travail ne déroge pas à cette règle. Plus particulièrement, nous nous proposons de définir un modèle d'interprétation de scène, représentée par une image, qui *simule* le comportement de l'-expert- humain face à une tâche d'interprétation de la scène. En effet, pour appréhender une scène, l'être humain ne s'adonne pas à la recherche des contours, ou à la classification des pixels. Il procède plutôt par une décomposition de l'image en régions grossières homogènes qu'il essaie de reconnaitre selon ce qu'il connait de ce type d'images. S'il n'arrive pas à reconnaitre toutes les régions alors il emploiera les informations sur les régions reconnues pour "qualifier" celles qui lui échappent. L'interprétation finale résulte de la comparaison des chacune des interprétations locales aux régions. Ainsi nous retenons de ce comportement de l'expert trois aspects importants du processus d'interprétation d'image que nous allons modéliser dans notre approche :

1. Le processus de segmentation n'est pas nécessairement séquentiel comme la plus part des techniques de segmentations qu'on rencontre, mais plutôt une suite de décisions pouvant remettre en cause leurs prédécesseurs. L'essentiel étant à la fin d'avoir la meilleure classification des régions. L'interprétation ne doit pas être limitée par la segmentation.
2. Le processus de caractérisation d'une zone d'intérêt n'est pas strictement monotone i.e. que l'expert peut aller d'une vue centrée sur la zone à une vue plus large incluant ses voisines pour ensuite retourner vers la vue contenant uniquement la zone et vice-versa.
3. Lors de la décision plusieurs sources d'informations sont sollicitées et fusionnées pour une meilleure certitude.

De cette analyse, nous proposons un modèle de segmentation d'images simulant ce comportement en se basant sur les informations du domaine des images. Ce modèle se repartit sur trois principales étapes qui sont la *focalisation* d'attention, la *propagation* des connaissances et la *fusion* d'information.

Les approches de segmentation à base de connaissances se divisent essentiellement en deux phases principales et complémentaires : l'acquisition des connaissances et le raisonnement sur ces connaissances. La première phase assure la collecte des connaissances à intégrer dans l'approche, mais aussi la formalisation et l'organisation de ces connaissances afin de leur garantir une utilisation optimale par la seconde phase de la segmentation. Dans l'approche que nous proposons, l'acquisition des connaissances se décompose en la définition du type des connaissances à intégrer, la représentation formelle de ces connaissances et leur organisation architecturale.
La phase de raisonnement regroupe toutes les étapes permettant de produire une image segmentée à partir d'une image donnée en utilisant les connaissances précédemment collectées. Il s'agit donc dans cette étape d'utiliser les connaissances pour manipuler les opérateurs du traitement d'images tels que le filtrage, la

classification ou la fusion des régions pour aboutir à la segmentation de l'image en entrée. Nous proposons une phase de raisonnement itérative qui se compose d'une étape de *focalisation d'attention*, suivie d'une étape de la *propagation et l'intégration des connaissances* pour finir par une étape de *fusion des régions* similaires. Cette modélisation proposée est résumée par la figure II.1.

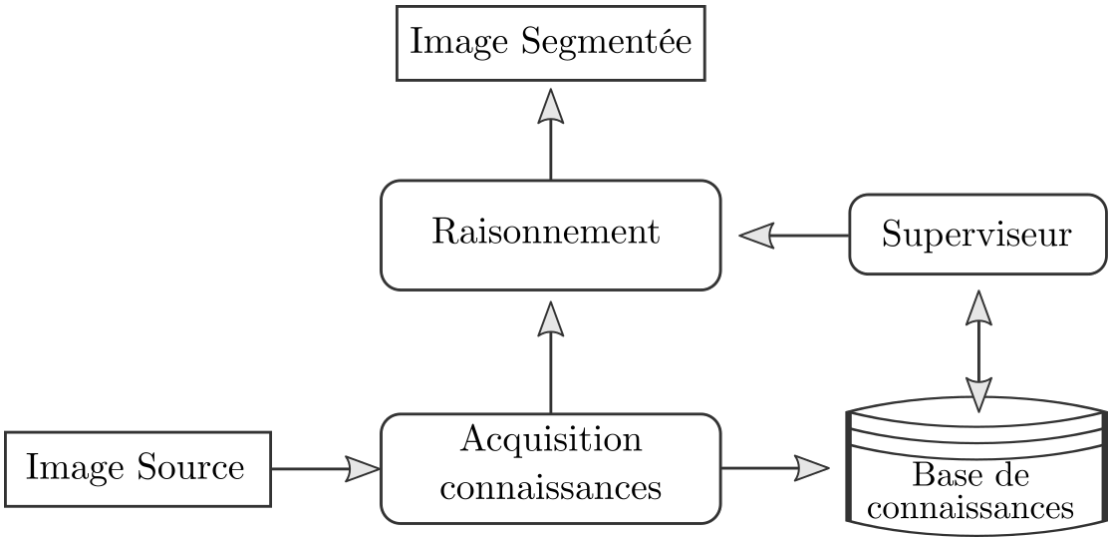


FIGURE II.1. – Schéma fonctionnel de l'approche de segmentation d'image proposée.

2. Acquisition des connaissances

Nous rappelons que notre travail s'intéresse à l'intégration des connaissances dans la segmentation et se focalise précisément sur les connaissances relatives au domaine d'images choisi sans considérer les connaissances sur les opérateurs du traitement d'images ou celles stratégiques (§ Chap.I, Sec.2.1.1). Cette catégorie de connaissance couvre toute connaissance se rapportant au domaine d'application de l'approche. Ce sont donc les connaissances permettant de caractériser le contenu des images à traiter.

L'acquisition des connaissances est une étape cruciale dans les approches de segmentation d'images à base de connaissances telle que celle proposée dans ce travail. D'une part, cette étape assure la collecte l'ensemble des connaissances sur lesquelles le raisonnement de l'approche sera basé : ceci constitue la base des connaissances de l'approche. D'autre part, la formalisation et la représentation de ces connaissances de façon à optimiser l'utilisation sont aussi réalisées durant cette étape. Conséquemment, le domaine de raisonnement de l'approche est limité par les connaissances fournies par l'acquisition des connaissances. Quelque soit l'efficacité du modèle de raisonnement implémenté dans l'approche, il ne saurait pas détecter un type d'objet qui ne soit présent dans la base des connaissances. Par ailleurs, le raisonnement est aussi fortement dépendent des paradigmes utilisés pour représenter les connaissances. Tout d'abord, il n'existe pas de formalisme qui soit parfait dans toutes les situations, mais aussi le formalisme utilisé définit entièrement les cas d'utilisation possibles des connaissances. Indépendamment du modèle de raisonnement ou même des connaissances considérées, une approche utilisant le formalisme des graphes pour représenter les connaissances manipulera très facilement les informations spatiales telles que le voisinage des objets. Le même traitement serait sans doute trop complexe à réaliser avec le formalisme des règles de production.

Dans le cadre de ce travail, nous distinguons deux catégories de connaissances à intégrer dans l'approche. Les connaissances *a priori* qui sont fournies par l'expert du domaine et les contraintes visuelles qui sont extraites automatiquement par l'approche à partir des images.

2.1. Connaissances *a priori*

Une connaissance *a priori* est toute connaissance décrivant de façon générique le type d'image considéré. Ces connaissances décrivent essentiellement les différentes classes d'objets observables dans les images et leurs positions relatives ou absolues. Ces connaissances sont en général fournies par les experts du domaine d'application sous forme de modèle générique du contenu des images. Ainsi, cette catégorie collectera toutes les connaissances disponibles afin de construire un modèle global du contenu typique des images du domaine d'application de l'approche. L'essentiel des connaissances *a priori* consiste en la caractérisation de l'ensemble des objets rencontrés dans le type d'image d'étude, aussi appelé *classes thématiques*. Cet ensemble est aussi appelé ensemble de classes modèles et est noté par $C = \{C_1, C_2, \ldots, C_K\}$.

2.1.1. Caractérisation des classes thématiques

Idéalement, chaque classe C_k est décrite par des patchs d'images fournissant un exemple de l'objet considéré. Dans ce cas, une procédure d'extraction des propriétés est réalisée afin de calculer à partir des patchs un vecteur de descripteurs $\mathcal{F}_{C_k} = (H_k, \mu_k, T_k, Pr_k)$ pour chaque classe C_k. Ce vecteur se compose de descripteurs de deux familles distinctes : les descripteurs numériques et les descripteurs thématiques, résumés dans le tableau II.1.

Descripteurs numériques :

Ils décrivent les patchs par leurs propriétés bas-niveaux issues directement du capteurs. Nous considérons principalement les descripteurs de couleurs (H_k et μ_k) et de texture (T_k).

— **Histogramme normalisé de couleur** H_k : il présente le pourcentage de chaque valeur de niveau de gris dans le patch.

$$H_k = \frac{1}{N} \times \{\, card(p \mid p = i) \,\}_{i \in [0,\ldots,255]} \tag{II.1}$$

— **Moments de couleur** μ_k : ils servent à décrire la couleur du patch avec des caractéristiques statistiques de la distribution des valeurs de la couleur. Nous considérons les trois premiers moments de couleur qui correspondent respectivement à la moyenne (m_0), l'écart-type (m_1) et l'asymétrie (m_2) de la distribution des valeurs des niveaux de gris.

$$\mu_k = \{\, m_i \,\}_{i \in [0,\ldots,2]} \tag{II.2}$$

— **Descripteurs de Haralick** T_k : il s'agit des 13 descripteurs de Haralick (Haralick *et al.*, 1973) qui sont calculés pour décrire la texture du patch. Ils contiennent entre autres l'entropie, l'asymétrie, la corrélation, la variance, etc.

$$T_k = \{\, T_k^i \,\}_{i \in [0,\ldots,12]} \tag{II.3}$$

Descripteurs thématiques :

Ils fournissent une description haut-niveau des patchs par rapport à ceux numériques. Nous proposons une description probabiliste Pr_k des classes. En effet, la théorie des probabilités permet la représentation d'informations incertaines et offre un cadre théorique bien établi pour la manipulation de ce type d'informations.

— **Probabilités des niveaux de gris** Pr_k : ils calculent la probabilité conditionnelle de réalisation de chaque niveau de gris x sachant la classe C_k. Nous choisissons d'estimer les probabilités via la méthode du noyau KDE (Kernel Density Estimation) (Parzen, 1962) en utilisant un noyau de forme gaussienne.

$$\begin{cases} Pr_k = & \{ Pr(x \mid C_k) \}_{\forall x \in [0,\ldots,255]} \\ = & \{ \texttt{KDE}(x) \} \end{cases} \tag{II.4}$$

Très fréquemment, les distributions de probabilités de plusieurs classes se chevauchent sur certaines plages de valeurs des niveaux de gris. Ce comportement entraine une indécision dans les prochaines étapes de classifications des superpixels. Pour corriger cela, nous introduisons une étape de réajustement des probabilités des classes thématiques comme suit.

$$Pr_k(x) = \begin{cases} Pr_k(x) - \texttt{max}\big(Pr(x) \setminus Pr_k(x)\big), & \texttt{si } Pr_k(x) = \texttt{max}(Pr(x)) \\ 0, & \texttt{sinon} \end{cases} \tag{II.5}$$

Les effets de ce comportement sont présentés dans le paragraphe 6.2. Dans plusieurs situations les patchs des objets ne sont pas disponibles, dans ce cas les objet peuvent être décrits par leur distribution de probabilités d'intensités fournies par les experts. Cette distribution correspond à l'histogramme des niveaux de gris des objets. La figure II.2 présente des exemples de patchs d'objets.

TABLE II.1. – Caractéristiques calculées pour chaque classe C_k. Les caractéristiques sont divisées en descripteurs numériques et thématiques.

Catégorie	**Descripteurs**	$\mathcal{F}_{C_k}$
Numérique	Histogramme de NG	H_k
	Moment de couleur	μ_k
	Descripteurs de Haralick	T_k
Thématique	Probabilités	Pr_k

2.1.2. Formalismes de représentation

Au vu des informations représentées par les connaissances *a priori*, nous proposons d'utiliser le formalisme orienté objet pour représenter les classes d'objets conjointement avec un graphe d'adjacence orienté pour exprimer les relations spatiales entre les classes. En effet, comme indiqué dans le chapitre I, le paradigme orienté objet permet de se focaliser sur la modularité et favorise le partage des connaissances sur leur classification (Crevier et Lepage, 1997). De plus, cette façon de faire assure une structuration rigide

pour une meilleure organisation des connaissances. A ce titre, il offre un très bon moyen pour modéliser les classes d'objet *a priori*.

Quant aux relations spatiales entre les classes d'objets, elles sont, par nature, d'ordre topologique. La structure de graphe constitue incontestablement le meilleur choix de représentation à cause de son expression naturelle de l'organisation topologique. De plus le nombre de classes est généralement assez petit, ce qui nous évite le problème de complexité dû à l'explosion du nombre des nuds, qui est souvent l'inconvénient majeur de ces structures de représentation.

Finalement, les connaissances *a priori* sont représentées par la combinaison des classes orientées objet et de la structure de graphe d'adjacence. Ceci nous conduit à une structure compacte de graphe d'adjacence dans laquelle chaque nud contient une classe du paradigme orienté objet. Les nuds représentent les classes modèles accompagnées par leurs caractéristiques tandis que le graphe reflète leurs relations spatiales dans une image typique du domaine d'application.

2.2. Contraintes visuelles

La deuxième catégorie de connaissances de notre approche cible les informations qui sont extraites de manière automatique à partir des images et qui peuvent permettre une meilleure caractérisation de leur contenu. Nous désignons par contrainte visuelle toute information qui décrit un trait particulier des images analysées extraites grâce à descripteurs invariants d'images. Ces descripteurs dépendent de chaque image et sont uniquement utilisés pour guider l'analyse de l'image courante. Dans beaucoup de cas, l'expression des contraintes visuelles à travers les connaissances *a priori* est complexe, voire impossible. La position absolue dans l'image d'une classe d'objet, la forme générique des objets exprimée par leur silhouette ou par leurs contours sont quelques exemples de ce type de connaissance. A l'évidence, ces connaissances trouvent leurs intérêts dans les images représentant des scènes fixes.

En pratique, nous définissons deux caractéristiques pour modéliser les contraintes visuelles sur les images : la silhouette $\mathcal{S}_{il}$ et les contours globaux $\mathcal{G}_C$. La première permet de capter l'organisation schématique des objets dans l'image. En effet, cette information permet de localiser les centres de gravité des objets dans l'image examinée et ainsi fournir des pseudo-régions à partir desquelles la constitution des objets peut amorcer. La deuxième contrainte visuelle modélise l'information complémentaire à la silhouette. En effet, les contours globaux représentent les bordures (frontières) qui viennent définir les

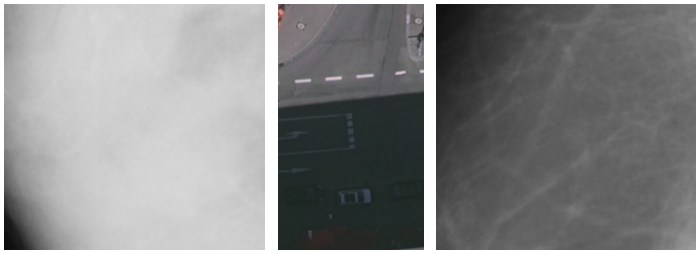


FIGURE II.2. – Quelques exemples de patchs de classes d'objets pour l'extraction des connaissances *a priori*.

limites précises des objets. Cependant, en tant que contrainte visuelle, les contours globaux extraites se limitent à ceux qui se démarquent nettement dans l'image. Les contours sont utilisés pour ajuster les similarités entre des régions voisines.

La silhouette correspond à une projection de l'image sur un espace de probabilités où les pixels les plus proches d'un contours sont les moins probables. Cela correspond à la même image dans laquelle les contours sont floutés. Ainsi, la silhouette est représentée par une matrice de valeurs réelles donnant pour chaque pixel sa proximité par rapport à un contour. De manière similaire, les valeurs des contours globaux sont contenus dans une matrice dans laquelle les contours sont représentés par la probabilité maximale de 1 pendant que le reste des pixels prend des valeurs inférieures selon leur proximité avec les contours jusqu'à atteindre la valeur nulle pour les pixels les plus éloignés.

Le tableau II.2 présente le récapitulatif des connaissances utilisées dans l'approche proposée.

TABLE II.2. – Tableau des connaissances et leur formalisme de représentation adopté.

Connaissance	**Descripteurs**	**Formalisme de représentation**
Connaissances *a priori*	Couleur Gradient Texture	Graphe d'adjacence & Classes orientées objet
Contraintes visuelles	Silhouette Contours globaux	Matrices de probabilités pixéliques

L'objet de la phase d'intégration itérative de connaissances est de changer les degrés d'appartenance des régions en tenant compte de leur voisinage. Cela permettra d'homogénéiser les étiquettes des régions et d'obtenir un étiquetage réaliste au sens des connaissances sur le type d'images considéré. Schématiquement, son algorithme se compose de trois phases principales qui sont itérées sur les masques M_k et les régions R_j jusqu'à la convergence de l'ensemble. Ces trois phases sont respectivement : la focalisation, l'intégration et la propagation des connaissances et enfin la fusion des régions.

2.3. Organisation architecturale

Une approche de segmentation d'image à base de connaissances dépend fortement des connaissances qu'elle utilise mais aussi de l'architecture de représentation de ces connaissances comme montré dans le chapitre « **Segmentation pour l'interprétation de scène** ». Rappelons que cette architecture définit le sens du flux des informations dans l'approche entre les différents niveaux des connaissances. Relativement à leur philosophie d'interprétation, beaucoup d'approche proposent une progression des connaissances des niveaux bas (pixel, primitives visuelles) vers les niveaux hauts (objets et scène) alors que d'autres approches optent plutôt pour le sens inverse. Aussi, les approches récentes, car beaucoup plus complexes, allient ces deux façons de faire pour donner une meilleure flexibilité d'utilisation des connaissances.

Il est clair que, dans le cadre des approches itératives balançant entre les différents niveaux de

connaissances entre le début et la fin de chaque itération, l'approche mixte constitue la meilleure option. Ainsi, dans notre cas le sens bottom-up est utilisé pour la caractérisation des régions en tenant compte des connaissances de l'approche tandis que le top-down nous permet de mettre à jour la description des régions "connaissant" les objets qui les entourent. En effet notre modèle de raisonnement à chaque itération se résume brièvement par l'identification des régions non consultées ensuite le réajustement des régions qui leurs sont connexes. Tel que présenté dans la figure II.1, les connaissances sont maintenues grâce à un système de Blackboard (Erman *et al.*, 1980) où elles sont exposées dans une base et accessibles pour toutes les étapes du processus de segmentation.

3. Décomposition en superpixels

Le superpixel est, depuis son avènement, de plus en plus utilisé dans beaucoup de travaux en Vision par Ordinateur. La littérature dans ce domaine présente un très grand nombre d'algorithmes de décomposition qui varient en fonction de la finalité des superpixels générés. Stutz *et al.* (2018) proposent une répartition des algorithmes de décomposition d'images en superpixels en sept catégories, comme suit :

Les algorithmes basés sur le watershed (Machairas *et al.*, 2015). Appelé aussi la Ligne de Partage des Eaux (LPE). Ils se distinguent généralement par la manière dont l'image est pré-traitée et par la manière dont les marqueurs sont définis. Le nombre de superpixels est déterminé par le nombre de marqueurs. Aussi, certains de ces algorithmes permettent aussi de contrôler la compacité.

Les algorithmes basés sur la densité (Comaniciu et Meer, 2002 ; Vedaldi et Soatto, 2008). Leur principe est d'utiliser une procédure itérative de recherche de mode pour trouver des modes dans l'espace colorimétrique ou d'intensité d'une image, pour localiser un maximum local d'une fonction de densité. Les pixels qui convergent vers le même mode forment le superpixel correspondant.En général, ces algorithmes produisent des superpixels de forme irrégulière sans dimensions uniformes mais ne permettent pas de contrôler le nombre de superpixels ou leur compacité.

Les algorithmes basés sur les graphes (Humayun *et al.*, 2015 ; Shi et Malik, 2000). Dans cette classe, les algorithmes transforment d'abord l'image en un graphe non dirigé pour ensuite partitionner ce graphe en fonction des pondérations des arêtes entre les nuds qui sont souvent calculées comme des différences de couleur. Ces algorithmes diffèrent dans l'approche de partitionnement où certains présentent une fusion ascendante des pixels en superpixels, tandis que d'autres utilisent des divisions successives de l'image.

Les algorithmes basés sur les contours (Buyssens *et al.*, 2014 ; Levinshtein *et al.*, 2009). Ces algorithmes représentent des superpixels en tant que contours évolutifs à partir des pixels initiaux. A noter que ces algorithmes nécessitent la sélection des pixels initiaux, qu'ils appellent germes.

Les algorithmes basés sur les chemins (Drucker et MacCormick, 2009 ; Fu *et al.*, 2014). Ces approches partitionnent une image en superpixels en connectant les points de départ via des chemins de pixels suivant des critères spécifiques. Le nombre de superpixels est facilement contrôlable, cependant, la compacité n'est généralement pas. Souvent, ces algorithmes utilisent des informations de contours en utilisant l'image des gradients ou toute autre approche de détection de contours.

Les algorithmes par clustering (Achanta *et al.*, 2012 ; Li et Chen, 2015 ; Neubert et Protzel, 2014). Ces algorithmes sont inspirés par des algorithmes de regroupement tels que k-means. Ils commencent par

une liste de pixels initiaux puis utilisant des informations de couleur et l'organisation spatiale ils assignent chaque pixels aux cluster qui lui est le plus proche. Intuitivement, le nombre de superpixels générés et leur compacité sont contrôlables. Bien que ces algorithmes soient itératifs, un post-traitement est nécessaire pour assurer la connectivité des superpixels.

Les algorithmes basés sur l'optimisation d'énergie (Van den Bergh *et al.*, 2015 ; Yao *et al.*, 2015). Ces algorithmes optimisent itérativement une énergie calculée sur les pixels de l'image. L'image est initialement partitionnée en une grille régulière pour constituer les superpixels initiaux. Par la suite, selon l'énergie des superpixels les pixels peuvent transiter entre les superpixels voisins. Le nombre de superpixels et la compacité sont contrôlables et les itérations peuvent généralement être interrompues à tout moment.

En général, la plupart des auteurs s'entendent sur les propriétés suivantes pour tout algorithme de décomposition en superpixels :

- **Partitionnement :** Les superpixels devraient définir une partition de l'image, c'est-à-dire que les superpixels doivent être disjoints et étiqueter chaque pixel.
- **Connectivité :** Les superpixels doivent représenter des ensembles connexes de pixels.
- **Adhésion aux contours :** Les superpixels devraient préserver les contours de l'image.
- **Compacité, régularité et continuité :** En l'absence des contours de l'image, les superpixels devraient être compacts, placées régulièrement et présenter des contours lisses.
- **Efficacité :** Les superpixels doivent être générés en un temps raisonnable.
- **Nombre contrôlable de superpixels :** Le nombre des superpixels générés doit être contrôlable par l'utilisateur.

Une étude plus complète sur les algorithmes de décomposition en superpixels peut être trouvée dans les travaux de Wang *et al.* (2017) et de Stutz *et al.* (2018).

3.1. Simple Linear Iterative Clustering (SLIC)

Le Simple Linear Iterative Clustering (SLIC) (Achanta *et al.*, 2010, 2012) est un algorithme basé sur le clustering et l'un des plus couramment utilisés pour générer des superpixels. Il offre une implémentation simple et fournit des superpixels compacts et presque uniformes. Les centres de cluster sont initialisés sur une grille uniforme et les pixels d'une fenêtre $2S \times 2S$ autour des clusters sont agrégés de manière itérative selon la métrique d (éq. II.6) définie dans un espace composé de cinq dimensions dont trois composantes couleur et deux composantes spatiales.

$$d = \sqrt{\left(\frac{d_c}{m}\right)^2 + \left(\frac{d_S}{S}\right)^2} \tag{II.6}$$

où $d_c = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2}$ est la distance entre les vecteurs de caractéristiques de couleur et $d_S = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$ est la distance entre les vecteurs de coordonnées spatiales du centre de cluster actuel C_k et le pixel voisin considéré i. m est un terme de pondération utilisé pour contrôler la régularité des superpixels générés et l'intervalle des centres initiaux de clusters $S = \sqrt{N/K}$ est la taille de la fenêtre de voisinage autour des centres de cluster ; N étant le nombre total de pixels dans l'image et K le nombre souhaité de superpixels de taille approximativement égale.

3.2. Caractérisation du superpixel

A ce niveau, l'image est représentée par l'ensemble des superpixels produits par l'algorithme de décomposition. Nous adoptons une représentation floue de l'image afin de réduire la dépendance par rapport aux valeurs numériques des (super)pixels de l'image mais aussi pour élargir les opérations applicables à l'image. En effet, l'immense majorité des méthodes de segmentation d'image opèrent sur les données de bas niveau qui sont les mesures numériques issues des capteurs (i.e. les intensités associées aux différents pixels) ou sur des grandeurs numériques calculées à partir de ces mesures. Cette forte dépendance, par rapport aux capteurs, constitue un inconvénient majeur de ces méthodes de segmentation car les régions détectées sont homogènes selon des critères appliqués aux données issues des capteurs et non pas selon des critères basés sur le contenu sémantique des régions constituantes. Ceci explique l'intérêt croissant pour l'intégration au sein des méthodes de segmentation d'image des connaissances a priori liées au contenu sémantique des images à segmenter.

En premier lieu, un superpixel P est décrit par ces caractéristiques notées par $\mathcal{F}_p$. Dans un souci de cohérence, les mêmes descripteurs que ceux des patchs de classes sont utilisés à cet effet (§ sec. 2.1). En second lieu, des descripteurs probabilistes sont ajoutées à ceux préalablement calculés sur la base de descripteurs thématiques (Pr) de $\mathcal{F}_p$.

3.2.1. Taux de présence thématique

Les taux de présence thématique calculent les proportions de chaque classe dans le superpixel considéré. Ainsi, pour chaque superpixel P les taux probabilistes $\tau(P)$ sont calculés comme suit :

$$\begin{cases} \tau(P) = \{\, \tau_k(P) \,\}_{k\in[1,\ldots,\mathrm{K}]} \text{ avec} \\ \tau_k(P) = Sim(Pr_p, Pr_k) / \sum_{j\in[1,\ldots,\mathrm{K}]} Sim(Pr_p, Pr_j) \end{cases} \tag{II.7}$$

La mesure de similarité $Sim(.,.)$ est définie au paragraphe 3.3. Pour chaque superpixel P de l'image, les taux probabilistes vérifient les conditions suivantes :

$$\forall P \in \mathcal{I}, \begin{cases} 0 \leqslant \tau(P) \leqslant 1 \\ \sum_{k\in[1,\ldots,\mathrm{K}]} \left(\tau_k(P) = 1\right) \end{cases} \tag{II.8}$$

En considérant l'ensemble des taux d'appartenance à une classe C_k comme un masque M_k de valeurs superposable à l'image des superpixels, on obtient à la sortie de cette caractérisation un ensemble de masques $\mathcal{M} = \{M_1, M_2, \ldots, M_K\}$ qui se substituera à l'image originale pour la suite du processus.

À partir de $\mathcal{F}_p$, nous définissons les éléments suivants pour un superpixel P :

— $\tau = (\tau_1, \tau_2, \ldots, \tau_K)^T$ par le vecteur des taux d'appartenance de P aux classes de C.

— $\hat{\tau}$ l'ensemble des taux d'appartenance de P trié selon l'ordre croissant.

— $\tilde{\tau} = max(\tau) = \hat{\tau}(1)$ le plus grand degré d'appartenance de P.

— $\tilde{C} = C_k \mid \tau_k = \tilde{\tau}$ la classe dominante selon les taux d'appartenance de P.

— $F = \{f_1, f_2, \ldots, f_{K-1}\}$ avec $f_i = \hat{\tau}(i+1) - \hat{\tau}(i)$ l'ensemble des écarts inter-classes de P ordonnés selon le taux de présence des classes. Ainsi, f_{K-1} donne l'écart de la classe $\tilde{C}$.

3.2.2. Voisinage des superpixels

En plus de leurs taux d'appartenance aux classes, les superpixels sont caractérisés par une description de leur voisinage $\mathcal{V}_P$ qui est constitué de l'ensemble des superpixels qui leur sont connexes. Nous adoptons le même formalisme de représentation de connaissances que celui utilisé pour les connaissances *a priori* (Sec. 2.1) i.e. un graphe dont les nuds représentent les superpixels et les arêtes expriment les relations de voisinages entre eux. De cette description, nous dégageons trois catégories de connaissances reparties sur deux niveaux de représentation des connaissances comme illustré par le schéma bloc de notre approche (figure II.3). Rappelons que les approches de traitements sémantiques d'images définissent quatre niveaux de représentation de connaissances suivants : le niveau *scène*, le niveau *objet*, le niveau *région* ou encore *primitives visuelles* et le niveau *pixel*. Dans notre travail nous nous sommes concentré sur les niveaux (*iii*) et (*iv*). D'autre part, nous divisons le niveau *région* en deux parties : la partie *région* qui regroupe les régions simples telles que définies par le processus de segmentation d'image i.e. un groupe de pixels voisins formant un ensemble homogènes selon un certain critère. La partie sémantique, appelée niveau *classe thématique* qui regroupe les régions auxquelles une étiquette est associée pour indiquer son appartenance à une classe modèle de la scène décrite.

Ainsi, les connaissances sur les superpixels sont exprimées par trois types de descripteurs. Les *descripteurs contextuelles* $\mathcal{V}_P$ qui reflètent les relations spatiales entre les superpixels. Les *descripteurs thématiques* Pr_P utilisés pour contenir les degrés d'appartenance des superpixels tandis que leurs caractéristiques bas niveau (i.e. densité d'intensités ou texture) sont spécifiés par les *descripteurs numériques* H_P, μ_P et T_P. La description finale du superpixel est alors $\mathcal{F} = (H_P, \mu_P, T_P, Pr_P, \mathcal{V}_P)^T$.

Après la décomposition en superpixels et la caractérisation de ces superpixels, les étapes suivantes de notre démarche utiliseront ces superpixels comme régions initiales pour produire une image segmentée. Pour atteindre cet objectif trois étapes -la **focalisation**, la **propagation** et la **fusion**- sont itérativement alternées pour raffiner leurs résultats respectifs et produire une segmentation précise de l'image.

3.3. Mesure de similarité

Une mesure de similarité est une fonction qui estime la ressemblance entre deux éléments d'un ensemble. La section 3.1 du chapitre « **Segmentation pour l'interprétation de scène** » présente un état-de-l'art des mesures de similarité entre régions dans un contexte de segmentation d'image par croissance de régions. Nous définissons la similarité Sim entre deux superpixels P et Q décrits respectivement par $X = (x_0, \ldots, x_K)$ et $Y = (y_0, \ldots, y_K)$, par l'équation II.9.

$$Sim(P,Q) = 1 - moy\big(\mu_{s_0}(d(x_0, y_0)), \ldots, \mu_{s_0}(d(x_K, y_K))\big) \tag{II.9}$$

Où $moy(\ldots)$ calcule la moyenne d'un ensemble de valeurs numériques. $d(x, y)$ correspond à la distance Euclidienne lorsque x et y sont des réels tandis que nous utilisons la distance de Bhattacharyya $d_{Bh}(p,q) = -ln\big(\sum_{x \in X} \sqrt{p(x)q(x)}\big)$ lorsqu'il s'agit des distributions de probabilités. Plusieurs formules existent dans la littérature pour capter la différence entre deux distributions de probabilités (Bhattacharyya, 1943, 1946 ; Kullback et Leibler, 1951 ; Shih et Cheng, 2005). Chaque mesure présente des propriétés spécifiques lui

permettant de convenir à des situations plus ou moins particulières. Dans le cadre de ce travail, la distance de Bhattacharyya (Bhattacharyya, 1943) donne les meilleurs résultats car celle-ci est en effet utilisée pour mesurer la séparabilité des classes dans un univers de probabilités. Elle s'exprime en fonction du coefficient du même auteur. La distance de Bhattacharyya augmente en fonction de la différence entre les écarts types. μ_{s_0} définit une contrainte d'acceptabilité d'un superpixel par :

$$\begin{cases} \mu_{s_0} : \mathbb{R}^+ \mapsto [0,1] & \\ \mu_{s_0}(x) = \frac{-1}{s_0}x - 1, & \texttt{si } x \leqslant s_0 \\ \mu_{s_0}(x) = 0, & \texttt{si } x > s_0 \end{cases} \tag{II.10}$$

La valeur s_0 indique la limite inférieure nécessaire pour vérifier la condition.

4. Raisonnement sur les connaissances

La propagation des connaissances ne peut se faire à partir de régions quelconques, car les informations relatives à certaines régions sont entachées d'incertitudes quant à leur appartenance à une classe précise. Il est par conséquent nécessaire de pouvoir choisir les régions jugées fiables pour propager les informations dont elles sont porteuses.

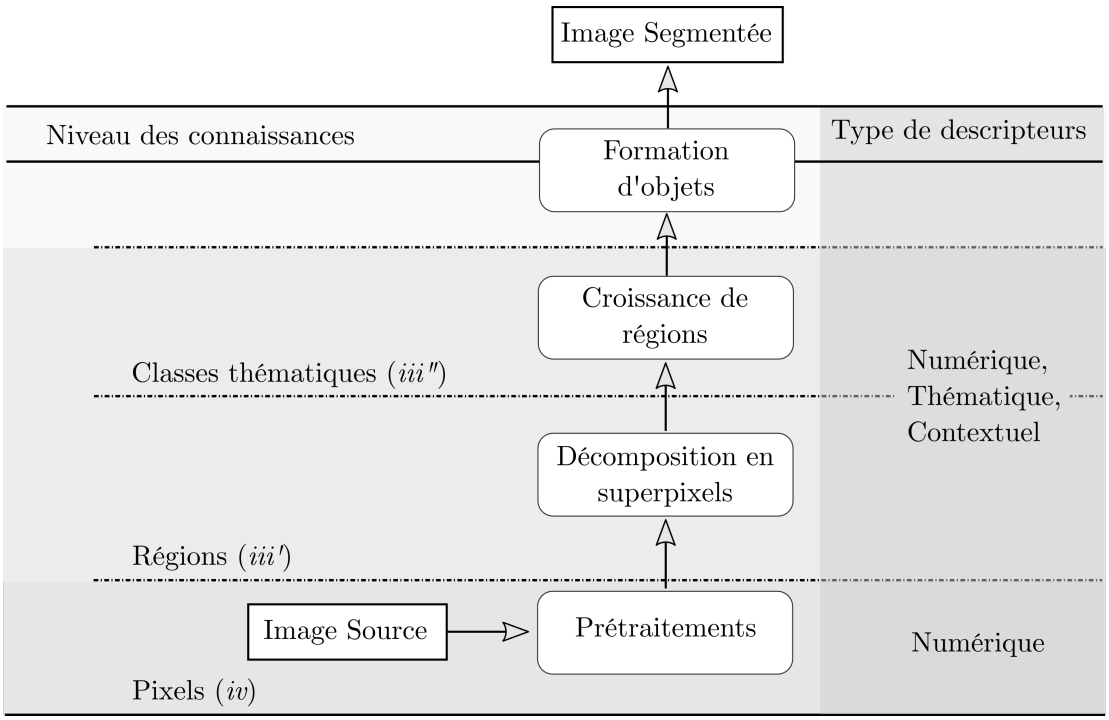


FIGURE II.3. – Positionnement de notre approche par rapport aux niveaux de représentation de connaissances dans une approche de vision par ordinateur. Notre approche se limite aux niveaux *pixel* et *primitives visuelles* car ceux-ci couvrent tous les types de connaissances utilisées dans notre approche.

4.1. Focalisation d'intérêt

Le principe de focalisation dans les approches de segmentation d'image se retrouve particulièrement dans les croissances de région avec germes ou **S**eeded **R**egion **G**rowing (SRG) (Adams et Bischof, 1994 ; Dreizin *et al.*, 2016 ; Huang *et al.*, 2018 ; Shih et Cheng, 2005 ; Thirumeni *et al.*, 2015). D'un point de vue simpliste, il est question de sélectionner des régions de l'image à partir desquelles les régions finales de la segmentation seront formées. Ces régions initiales correspondent aux différentes parties homogènes de l'image qui seront utilisées afin de guider le processus de formation des régions finales. Tandis que certains auteurs laissent le choix de ces régions à l'utilisateur (qui est le cas des segmentations interactives), d'autres (Huang *et al.*, 2018 ; Shih et Cheng, 2005) proposent une génération automatique de ces régions. Notre approche adopte cette dernière façon de faire. Ainsi dans le cadre de ce travail, la focalisation consiste à choisir parmi l'ensemble des superpixels $\mathcal{P}$ ceux qui sont assez bien qualifiés i.e. dont l'appartenance à au moins une des classes C_k est "certaine". Cela implique de définir un seuil de certitude aux taux d'appartenance ainsi qu'une heuristique de choix de régions adéquates. La focalisation est donc une fonction qui peut se définir comme suit :

$$\begin{cases} f : \mathcal{P} \times \mathcal{M} \rightarrow \{\texttt{vrai}, \texttt{faux}\} \\ f(\mathcal{P}_j, \mathcal{M}) = c \end{cases} \tag{II.11}$$

Avec $\mathcal{P}$ et $\mathcal{M}$ respectivement l'ensemble des superpixels et celui de leurs masques d'appartenance.

4.1.1. Méta-classification des superpixels

Les masques $\mathcal{M}$ des taux d'appartenance des superpixels contiennent la description thématique des superpixels. Dans l'objectif de choisir les superpixels germes, nous proposons une classification des superpixels selon les valeurs de leur taux d'appartenance. Ceci offre une qualification symbolique de l'information au niveau du superpixel et permet de juger de sa pertinence. La notion de méta-classification se réfère donc à la classification des taux d'appartenance qui sont eux-mêmes issus du processus de classification probabiliste.
Dans ce deuxième niveau de classification, les superpixels sont divisés en trois groupes, illustrés par des exemples dans la figure II.4 :

— les superpixels *reconnus* $\mathcal{P}_2$: ils représentent les cas où les taux d'appartenance exhibent une nette domination d'une seule classe par rapport aux restes.

— les superpixels *incertains* $\mathcal{P}_1$: ce groupe contient les superpixels présentant une classe dominante mais qui ne possède pas suffisamment d'avance sur les autres classes.

— les superpixels *ambigus* $\mathcal{P}_0$: il s'agit des superpixels dont la distribution des taux d'appartenance est stationnaire i.e. toutes les classes ont à peu près les mêmes poids.

Les ensembles de superpixels issus de la méta-classification sont définis par :

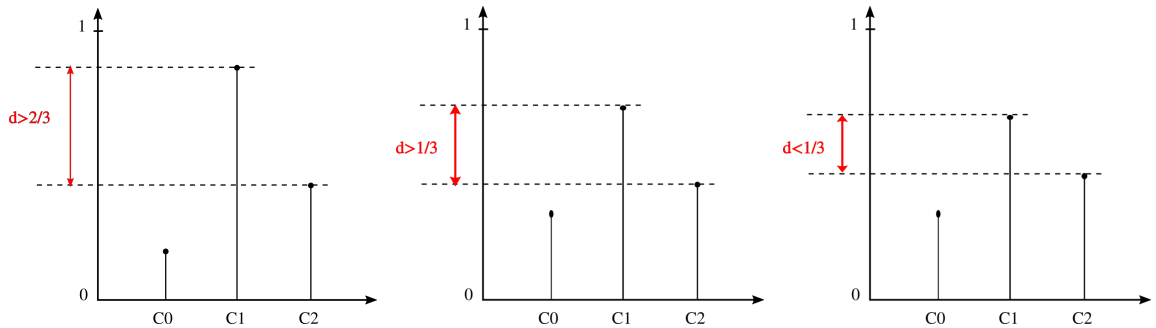
$$\mathcal{P}_\alpha = \{P \in \mathcal{P} \mid \quad f_{K-1} \geqslant \frac{\alpha}{K}\} \quad \forall\, \alpha \in \{\texttt{0},\texttt{1},\texttt{2}\} \tag{II.12}$$

Avec $\mathcal{P}$ l'ensemble des superpixels de l'image et K le nombre des classes d'objets prédéfinies. Rappelons que f_{K-1} donne l'écart de valeur du taux d'appartenance de P à sa classe dominante et celle qui la précède.

4.1.2. Sélection des germes

Compte-tenu des définitions précédentes, P est un superpixel germe lorsque $f_{K-1} \geqslant \frac{2}{K}$; ce qui correspond aux superpixels *reconnus*. La figure II.4 illustre le principe de sélection de germes proposé. L'ensemble des superpixels germes $\tilde{\mathcal{P}}$ est un sous-ensemble de $\mathcal{P}$ qui se caractérise par l'équation suivante :

$$\tilde{\mathcal{P}} = \{P \mid f_{K-1} \geqslant \frac{2}{K};\ P \in \mathcal{P}\} \tag{II.13}$$



(a) Superpixel reconnus (b) Superpixel incertains (c) Superpixel ambigus

FIGURE II.4. – Schéma illustratif des trois classes de la méta-classification. La meilleure configuration (II.4a), suivie de celles des superpixels à qualification moyenne (II.4b) et enfin les superpixels dont aucune classe ne se démarque des autres (II.4c).

Dans le principe de sélection de germes tel que présenté ci-haut, le superpixel est choisi par analyse de son information intrinsèque à savoir ses taux d'appartenance. Nous proposons d'ajouter une seconde analyse qui portera sur la totalité de l'information sur l'image en cours de traitement. Cela permet de juger la qualité de l'information du superpixel par rapport au reste de l'image. Ainsi, nous définissons pour chaque classe C_k l'ensemble des taux d'appartenance des superpixels à C_k par $D_k = \{d_i\}_{i \in [1,\ldots,N]}$. Partant de ceci, les superpixels germes de C_k, sont alors ceux qui sont supérieurs au dernier quartile de D_k, c'est-à-dire :

$$\tilde{\mathcal{P}} = \{P \mid f_{K-1} \geqslant \frac{2}{K} \text{ ou } \tau_k(P) \geqslant q_3;\ P \in \mathcal{P}\} \tag{II.14}$$

Cette définition nous permet d'avoir des superpixels germes robustes tout en maintenant une vision dynamique centrée sur l'image considérée. En particulier cela nous garantit d'avoir des germes pour chaque classe thématique quelque soit l'image d'entrée.

4.2. Propagation et intégration des connaissances

La sélection des superpixels germes permet de détecter les superpixels foyers de l'image à partir desquels le processus de segmentation itératif sera activé. La propagation et l'intégration consiste, tout d'abord, à la diffusion de l'information contenue au niveau du superpixel germe vers les superpixels non-germes qui lui sont connexes sous forme de connaissances utiles. Par la suite, ces connaissances sont fusionnées avec l'information au niveau de ces superpixels voisins pour améliorer leur caractérisation. Ainsi grâce aux connaissances *a priori*, la confirmation d'une classe d'objet C_k au niveau d'un superpixel P peut permettre de renforcer ou d'atténuer la présence d'autres classes dans la caractérisation des superpixels connexes à P. À l'inverse de la technique de croissance de régions qui change l'étiquette des pixels à chaque itération,

notre approche permet d'étaler la transition d'étiquette d'un superpixel sur plusieurs itérations. Cela permet une transition avertie et plus robuste.

Concrètement, le processus de propagation de connaissances (résumé dans l'algorithme II.1 et illustré par la figure II.5) à partir de l'ensemble des régions initiales propage les connaissances entre les superpixels voisins pour aboutir à un nouvel ensemble de régions. Ce processus présente deux grandes étapes : la croissance numérique et la croissance thématique. Le résultat de ces deux croissances est fusionné par une étape de validation qui génère une carte de régions final de l'itération courante à partir des cartes issues de deux croissances. Chaque phase de croissance prend en entrée un ensemble de régions qu'elle met à jour en fonction des propriétés spécifiques des superpixels.

Algorithme II.1 : Propagation de connaissances

Données : $\mathcal{I}, \mathcal{M}_\mathcal{R}$; ▷ $\mathcal{I}$:Image source, $\mathcal{M}_\mathcal{R}$:Ensemble des régions
Résultat : $\mathcal{M}_\mathcal{R}$
1 **début**
2 $\mathcal{M}_\mathcal{R}$:= DecompositionSuperpixels($\mathcal{I}$) ;
3 arret := Faux ;
4 **répéter**
5 $\mathcal{M}_N$:= CroissanceNumerique($\mathcal{M}_\mathcal{R}$) ;
6 $\mathcal{M}_T$:= CroissanceThematique($\mathcal{M}_\mathcal{R}$) ;
7 $\mathcal{M}'_\mathcal{R}$:= Validation($\mathcal{M}_N, \mathcal{M}_T$) ;
8 arret := Convergence($\mathcal{M}_\mathcal{R}, \mathcal{M}'_\mathcal{R}$) ;
9 $\mathcal{M}_\mathcal{R} := \mathcal{M}'_\mathcal{R}$;
10 **jusqu'à** *arret* = *Vrai* ;

4.2.1. Croissance numérique

En l'absence des connaissances *a priori* sur les images à traiter, notre approche produit une segmentation "*numérique*" des images. Cette segmentation est dite numérique car le calcul de la similarité entre deux superpixels voisins P et Q est réalisée sur les descripteurs bas-niveaux des superpixels, (H_P, μ_P, T_P) et (H_Q, μ_Q, T_Q). Lorsque la similarité entre P et Q est supérieure à un seuil $\mathcal{S}_0$ alors P et Q sont assignés à la même région R.

NB : Sachant que cette étape se limite au niveau numérique des connaissances, aucune propriété ne permet de distinguer les superpixels entre eux. Ainsi, la sélection des germes n'a pas de sens particulier mais renvoie tous les superpixels.

4.2.2. Croissance thématique

Dans cette étape, le processus d'intégration et de propagation pour un superpixel P non-germe se résume en deux opérations : la détermination de la configuration de mise-à-jour (définies à partir des connaissances *a priori*, tableau II.3) ensuite l'application de la mise-à-jour en fonction de la configuration choisie. Premièrement les caractéristiques de P sont modifiées en fonction de celles des superpixels qui l'entourent (aussi appelés son contexte local). Cette transformation a pour but d'assurer la cohérence de P avec son voisinage.

Deuxièmement, les informations issues de cette modification seront diffusées aux superpixels connexes à P afin de répandre leurs effets. Cette étape permet principalement de propager les informations des superpixels les mieux classifiés vers les plus "*incertaines*".

Configurations de mise-à-jour des régions

La mise à jour se base sur la méta-classification des superpixels selon leurs taux d'appartenance ; à savoir les superpixels *reconnus*, les superpixels *incertains* et les superpixels *ambigus*. Les configurations de mise-à-jour présentent tous les cas possibles de dispositions spatiales des objets (représentant les classes thématiques) pouvant se trouvées dans le type d'image traité. Nous nous limitons aux configurations entre deux classes et proposons, pour tout couple de classes voisines, deux dispositions spatiales possibles : le voisinage et l'inclusion. Tandis que dans le premier cas les deux classes ne partagent qu'une partie de leur bordure, dans le second cas l'une des classes est entièrement incluse dans l'autre.
Dans l'espace des superpixels ces formulations se rapportent sur les classes dominantes (§ **Sec.** 4.1, §. **Sélection des germes**) des superpixels considérés. Par ailleurs, pour un superpixel non-germe P, chaque configuration est associée à une règle de mise-à-jour des descripteurs du superpixel en fonction de son contexte local $\mathcal{V}_P$, telles que définies dans le tableau II.3.

TABLE II.3. – Tableau des configurations et leur règles de mise-à-jour du superpixel P tenant compte de son contexte local $\mathcal{V}_P$.

Configuration	**Règles de Mise-à-jour**
$\mathcal{V}_P$ homogène, "physiquement [1]" similaire à P	Augmenter le taux de P représentant la classe thématique de $\mathcal{V}_P$.
$\mathcal{V}_P$ homogène, "physiquement" non similaire à P et **tous** les superpixels de $\mathcal{V}_P$ sont *reconnus*	Augmenter le taux des classes pouvant être incluses dans la classe de $\mathcal{V}_P$.
$\mathcal{V}_P$ homogène, "physiquement" non similaire à P, **certains** superpixels de $\mathcal{V}_P$ sont *reconnus*	Augmenter le taux des classes pouvant être voisines de la classe de $\mathcal{V}_P$.
$\mathcal{V}_P$ hétérogène	Augmenter le taux des classes formant une configuration valide avec $\mathcal{V}_P$.

NB : Dans certaines configurations il existe plusieurs classes candidates pour l'augmentation des taux d'appartenance cela est géré de deux manières suivantes qui sont mutuellement exclusives :

— Soit en considérant toutes les classes candidates donc augmenter les taux de toutes ces classes au niveau de P.

— Soit en choisissant une parmi toutes ces classes. Dans ce cas la classe qui a le plus grand taux d'appartenance est choisie.

Fonctions de mise à jour

Pour chacune des configurations précédentes, la mise à jour est effectuée en fonction de la différence d_P entre le superpixel et ses voisins. Cette dernière distance est égale à la moyenne pondérée des distances $d(\cdot)$

1. la similarité entre le superpixel et son voisinage est calculée à partir de leurs descripteurs bas-niveau.

du superpixel avec chacun de ses voisins pris à part.

$$d_P = \frac{1}{L}\sum_{l=1}^{L} d_{Bh}(P, V_l) \tag{II.15}$$

Où V_l est le $l^{ième}$ superpixel dans $\mathcal{V}_P$.

La mise à jour des taux d'un superpixel consiste à partitionner les taux d'appartenance de ce dernier en deux parties ∇ et $\bar{\nabla}$ (définies par les Règles de Mise-à-jour de la table II.3) et leur appliquer un des deux traitements ci-après :

- **Conforter sa classification :** Augmenter les taux d'appartenance de ∇ d'une valeur de $\epsilon = \frac{\sum_{j\in\nabla}(\tau_j)}{|\nabla|} \times \frac{1-d_P}{K}$ et de réduire ceux de $\bar{\nabla}$ de $\frac{\epsilon}{K-|\nabla|}$ ou bien
- **Corriger sa classification :** Réduire les taux d'appartenance de ∇ de ϵ et augmenter ceux de $\bar{\nabla}$ de la même valeur.

$|\cdot|$ désigne le nombre d'éléments dans un ensemble.
Dans les deux cas, la valeur à ajouter ou à réduire est définie en fonction de la différence entre le superpixel et son contexte local. D'ailleurs, la quantité ϵ représente la différence entre le superpixel P et son contexte local pondéré par la moyenne des classes à augmenter l'appartenance. Il est ainsi évident que ce processus permettra la convergence vers une cohérence du superpixel avec son voisinage.
Ainsi, l'équation II.23 présente les nouvelles valeurs du vecteur d'appartenance du superpixel, après mise à jour.

$$\begin{bmatrix} \tau_1 \\ \vdots \\ \tau_k \\ \vdots \\ \tau_K \end{bmatrix} = \begin{bmatrix} \tau_1 - \frac{\epsilon}{K-|\nabla|} \\ \vdots \\ \tau_k + \epsilon \\ \vdots \\ \tau_K - \frac{\epsilon}{K-|\nabla|} \end{bmatrix} \tag{II.16}$$

Avec C_k parcourant l'ensemble des classes appartenance à ∇.

4.2.3. Validation de la propagation

Cette étape permet de combiner et valider les résultats de la croissance numérique et thématique. Elle produit en sortie un ensemble de régions qui cumule la propagation numérique et thématique et qui constitue l'entrée de l'itération suivante. Elle se garantit deux conditions principales : le respect des contours globaux et celui de la silhouette. Ainsi, pour deux maps $\mathcal{M}_N$ et $\mathcal{M}_T$ la première condition est assurée par $\mathcal{M}_{G1} = \mathcal{M}_N \otimes \mathcal{G}_C$ et $\mathcal{M}_{G2} = \mathcal{M}_T \otimes \mathcal{G}_C$. Tandis que les contours globaux sont utilisés pour diviser des superpixels voisins regroupés à tord, la silhouette permet de renforcer le regroupement par $\mathcal{M}_{S1} = \mathcal{M}_N \otimes \mathcal{S}_{il}$ et $\mathcal{M}_{S2} = \mathcal{M}_T \otimes \mathcal{S}_{il}$. Ainsi, nous avons le map résultat $\mathcal{M}$ comme suit :

$$\begin{aligned} \mathcal{M} =& \mathcal{M}_{R1} \otimes \mathcal{M}_{R2} \\ & \texttt{avec} \quad \mathcal{M}_{R1} = \mathcal{M}_{G1} \otimes \mathcal{M}_{G2} \quad \texttt{et} \quad \mathcal{M}_{R2} = \mathcal{M}_{S1} \otimes \mathcal{M}_{S2} \end{aligned} \tag{II.17}$$

Les matrices $\mathcal{G}_C$ et $\mathcal{S}_{il}$ représentent respectivement l'image des contours globaux de l'image et celle de la silhouette. L'opérateur $\otimes$ calcule le produit de deux matrices. Dans le cas des contours globaux, il divise

deux superpixels qui sont traversés par un contour. Dans le cas de la silhouette, il sépare les superpixels qui n'appartiennent pas à la même région.

4.3. Fusion : Segmentation par agrégations

Les itérations de la phase d'intégration de connaissances accumulent au niveau des superpixels des informations qui, au fur et à mesure, leur permettent de changer leurs descripteurs pour tendre vers une homogénéité avec leur entourage. Après l'actualisation de leurs connaissances, certains superpixels connexes atteignent un seuil de ressemblance leur permettant de se fusionner. Ainsi, la phase de fusion, qui permet de regrouper deux superpixels en un seul, n'est pas systématique mais n'intervient que lorsque ses conditions sont remplies. Ces dernières conditions sont définies à travers des heuristiques sur les connaissances relatives aux superpixels.

Cette phase représente une défuzzification des cartes de probabilités suivie d'un regroupement des régions ayant la même étiquette. Ainsi, en se basant sur la méta-classification, on peut produire deux segmentations de l'image : soit la segmentation contenant uniquement les superpixels *reconnus*, soit la segmentation contenant les superpixels *reconnus* et les superpixels *incertains*.
La figure II.5 un schéma qui résume les principales étapes du processus de raisonnement sur les connaissances.

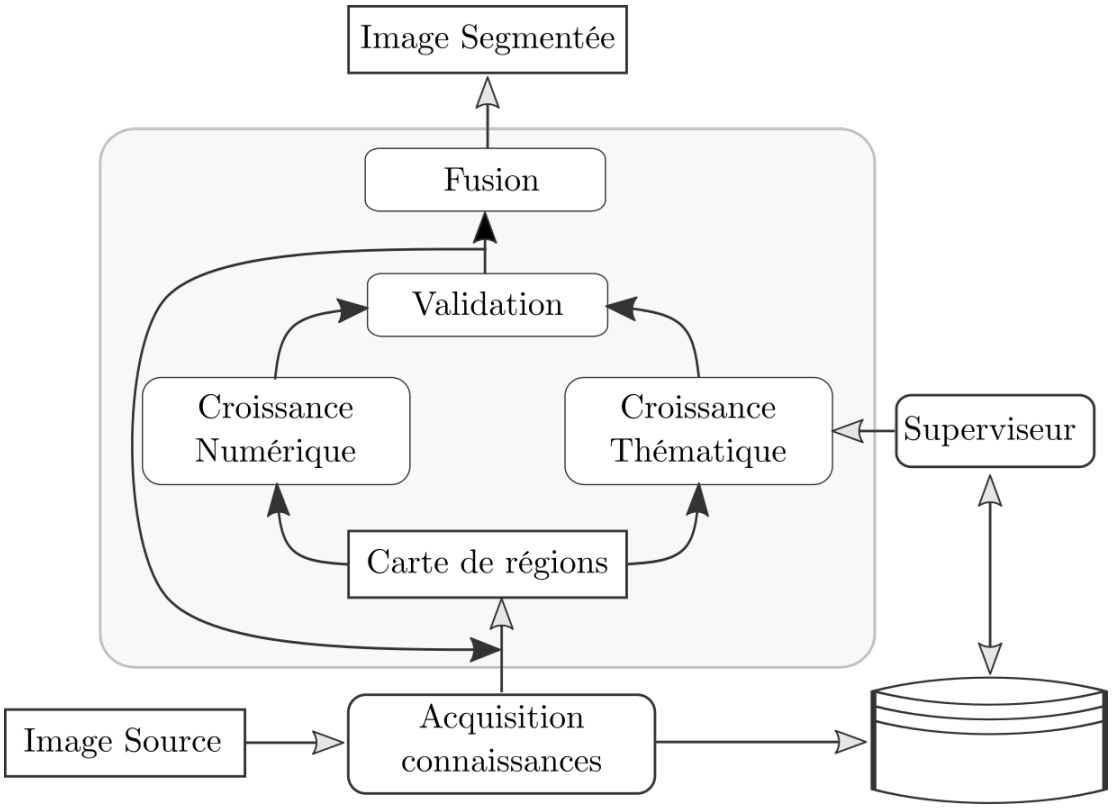


FIGURE II.5. – Récapitulatif des étapes du raisonnement sur les connaissances.

5. Raisonnement possibiliste

La théorie a été introduite par Zadeh en 1978 comme une extension de la théorie des ensembles flous et de la logique floue (Zadeh, 1978). Elle a ensuite été développée et décrite en profondeur par Dubois et Prade (Dubois et Prade, 1988). En effet, elle ajoute aux avantages de la théorie des ensembles flous, la gestion de l'ambiguïté de l'information, par exemple un niveau de gris d'une région qui est décrit comme "*sombre*" ou "*clair*" au lieu d'une valeur numérique donnée.

La mesure de possibilité, dans un univers fini Ω, est une fonction π à valeurs dans `[0,1]` et qui satisfait les propriétés suivantes :

$$\begin{cases} \pi(\emptyset) = 0 \text{ et } \pi(\Omega) = 1 \\ \pi(A \cup B) = max\big(\pi(A), \pi(B)\big), \quad \forall \quad A, B \subseteq \Omega \\ \pi(A \cap B) = min\big(\pi(A), \pi(B)\big), \quad \forall \quad A, B \subseteq \Omega \end{cases} \tag{II.18}$$

Lorsque $\pi(A) = 0$ alors A est dit évènement impossible, tandis que lorsque $\pi(A) = 1$ alors A est dit évènement normal. La mesure de possibilité d'un évènement A peut être complémentée par celle de sa nécessité N par :

$$N(A) = 1 - \pi(A) \tag{II.19}$$

Dans la théorie des possibilités, l'incertitude liée à la réalisation d'un événement $A \in \Omega$ est exprimée par sa possibilité et sa nécessité (Dubois, 1980). L'utilisation de ces deux mesures dans la théorie des possibilités permet d'encadrer la probabilité de réalisation de cet événement A tel que $Pr(A) \leqslant \pi(A) \quad \forall A \subseteq \Omega$.

Étant donné un univers $\Omega = \{x_1, x_2, \ldots, x_N\}$, la valeur $\pi(x_n)$ encapsule nos connaissances liées à l'occurrence du singleton x_n. En d'autre termes, $\pi(x_n)$ représente dans quelle mesure il est possible que le singleton x_n soit l'unique singleton qui s'est produit. Dans ce contexte, deux cas extrêmes des connaissances sont donnés :

— Connaissance complète : $\exists! x_n \in \Omega$, $\pi(x_n) = 1$ et $\pi(x_m) = 0$, $\forall\, x_m \in \Omega$, $x_m \neq x_n$;

— Ignorance totale : $\forall x_n \Omega, \pi(x_n) = 1$. Tous les singletons sont considérés comme tout à fait possibles.

La Connaissance complète exprime la situation dans laquelle un seul singleton possède la possibilité maximale et les autres évènements sont impossibles. Dans un cas d'Ignorance totale, tous les évènements ont la même valeur de possibilité maximale.

5.1. Caractérisation des classes et superpixels

Les superpixels sont caractérisés par une distribution de possibilités en plus de celle des probabilités d'appartenance. Plusieurs approches ont été proposées pour transformer une distribution de probabilités en une distribution de possibilités (Dubois et Prade, 1983 ; Zadeh, 1978). Dans ce travail, les distributions de possibilités sont générées à partir des probabilités en utilisant la transformation de Dubois-Prade

$P_r - \pi$ (Dubois et Prade, 1983, 1988) (Équation II.20).

$$\begin{cases} \pi_k = \{\, \pi(x \mid C_k)\}_{\forall x \in [0,\ldots,255]} \\ \pi(x \mid C_k) = Dubois - Prade_{Tansf}(\, Pr(x \mid C_k)\,) \\ \qquad\qquad = \sum_{j}^{K} min(\, Pr(x \mid C_k),\, Pr(x \mid C_j)\,) \end{cases} \tag{II.20}$$

Cette formule est dite "symétrique" puisqu'elle permet de passer des probabilités aux possibilités et inversement. Elle vérifie aussi les deux propriétés définies par ses auteurs comme suit :

— **Principe de cohérence** défini par : $N(A) \leqslant Pr(A) \leqslant \Pi(A), \quad A \subseteq \Omega$.

— **Principe de préservation** défini par : $Pr(A) < Pr(B) \Longleftrightarrow \Pi(A) < \Pi(B), \quad A, B \subseteq \Omega$.

Où $\Pi(A) = max_{x \in A}\big(\pi(x)\big)$, $Pr(A) = \sum_{x \in A} Pr(x)$ et $N(A) = 1 - \Pi(\bar{A})$.

Ainsi, la description des classes et des superpixels est étendue par l'ajout de descripteurs possibilistes des niveaux de gris π_k. Ils estiment la possibilité de réalisation chaque niveau de gris x sachant la classe C_k.

De façon similaire aux taux probabilistes, les taux possibilistes τ' sont calculés pour chaque superpixel par :

$$\begin{cases} \tau'(P) = \{\, \tau'_k(P)\,\}_{k \in [1,\ldots,\mathrm{K}]} \text{ avec} \\ \tau'_k(P) = Sim(\pi_p, \pi_k) / \sum_{j \in [1,\ldots,\mathrm{K}]} \big(Sim(\pi_p, \pi_j)\big) \end{cases} \tag{II.21}$$

Avec $Sim(X, Y)$ la mesure de similarité entre deux distributions de réels qui est définie par l'équation II.9 et π_p la distribution des possibilités du superpixel P.

5.2. Segmentation possibiliste

L'utilisation des distributions de possibilités ne change pas l'architecture principale de l'approche. Les changements apportés concernent essentiellement les équations manipulant les données. Ainsi, l'équation de sélection des régions germe II.14 devient :

$$\tilde{\mathcal{P}} = \{P \mid f_{K-1} \geqslant \frac{2}{K} \text{ ou } \tau'_k(P) \geqslant \mathrm{q}_3;\ P \in \mathcal{P}\} \tag{II.22}$$

Dans laquelle, f_{K-1} représente l'écart entre les deux plus grandes valeurs des taux de possibilités d'appartenance aux classes C_k et q_3 est le dernier quartile de l'ensemble des taux de possibilités d'une classe C_k. Les formules de l'étape de croissance thématique sont aussi mises-à-jour comme suit :

$$\begin{bmatrix} \tau'_1 \\ \vdots \\ \tau'_k \\ \vdots \\ \tau'_K \end{bmatrix} = \begin{bmatrix} \tau'_1 - \frac{\epsilon}{K - |\nabla|} \\ \vdots \\ \tau'_k + \epsilon \\ \vdots \\ \tau'_K - \frac{\epsilon}{K - |\nabla|} \end{bmatrix} \tag{II.23}$$

où $\epsilon = \frac{\sum_{j \in \nabla}(\tau'_j)}{|\nabla|} \times \frac{1 - d_P}{K}$. ∇ est le sous-ensemble des taux de possibilités de P dont la valeur va être augmentée pour la propagation.

6. Évaluation expérimentale

Nous proposons dans ce chapitre une approche de segmentation d'images par croissance de régions dont les principales contributions sont faites dans la sélection des germes et la propagation des connaissances des régions. Cette section expérimentale présente une évaluation de ces contributions par des critères établis dans la littérature.

6.1. Méthodes et critères d'évaluation

L'évaluation d'une approche de segmentation est une tâche fondamentalement difficile, voire "subjective" (Zhang *et al.*, 2008), malgré les nombreux travaux de recherche dans ce contexte. Par conséquent, cette évaluation ne permet pas une bonne appréhension de la séquence des étapes de segmentation. Étant donné que ce chapitre du manuscrit présente des contributions sur la focalisation et la propagation des connaissances dans une approche de segmentation sémantique, les méthodes d'évaluation non supervisées nous semblent les mieux adaptées à l'évaluation des résultats.

Dans les méthodes non-supervisées d'évaluation de segmentation une bonne évaluation optimise l'uniformité des pixels dans chaque région et minimise l'uniformité entre les régions. En pratique, ces méthodes génèrent une métrique d'évaluation de la segmentation, sans nécessiter une image de référence pour toute image quelconque. Trois méthodes sont sélectionnées pour évaluer nos résultats parmi les nombreuses approches qui ont été proposées dans la littérature (Borsotti *et al.*, 1998 ; Rosenberger et Chehdi, 2000 ; Zhang *et al.*, 2003)
Pour une image segmentée I, notons par :

- N_R le nombre de régions dans I
- $|R_i|$ le nombre de pixels dans la région R_i
- $I(p)$ la valeur du pixel p dans I
- $|x|$ la valeur absolue de x

Ainsi, nous avons :

1. Q (Borsotti *et al.*, 1998) : cette mesure se base sur le nombre, la surface et la variance de la surface des régions dans I pour produire une évaluation de la qualité de la segmentation. Elle formulée par :

$$Q(I) = \frac{1}{1000 * S_I}\sqrt{N_R}\sum_{i=1}^{N_R}\left[\frac{e_i^2}{1 + log|R_i|} + \left(\frac{N(|R_i|)}{|R_i|}\right)^2\right] \tag{II.24}$$

$$e_i^2 = \sum_{x\in\{r,g,b\}}\sum_{p\in R_i}\left(C_x(p) - \sum_{p\in R_i} C_x(p)\right)^2$$

   où $C_x(p)$ désigne la valeur du pixel p dans la canal x de l'image et e_i^2 dénote l'erreur quadratique couleur de la région R_i.

2. *Ros* (Rosenberger et Chehdi, 2000) : l'originalité de cette méthode réside dans son calcul adaptatif en fonction du type de région (uniforme ou texturé). Le critère calculé permet d'estimer l'homogénéité intra-régions et la disparité inter-régions. Ce critère quantifie la qualité d'un résultat de segmentation

comme suit :

$$Ros(I) = \frac{1 + \overline{D}(I) - \underline{D}(I)}{1} \tag{II.25}$$

$$\overline{D}(I) = \frac{1}{N_R} \sum_{i=1}^{N_R} \frac{|R_i|}{|I|} \overline{D}(R_i)$$

$$\underline{D}(I) = \frac{1}{N_R} \sum_{i=1}^{N_R} \frac{|R_i|}{|I|} \underline{D}(R_i)$$

où $\overline{D}(I)$ correspond à la disparité totale inter-régions qui quantifie la disparité de chaque région voisine de l'image I. La disparité totale intra-région, notée $\underline{D}(I)$, calcule l'homogénéité de chaque région de l'image I. $\overline{D}(R_i)$ et $\underline{D}(R_i)$ désignent respectivement la disparité inter- et intra-régions de la région R_i.

3. E (Zhang *et al.*, 2003) : ces auteurs utilisent l'entropie pour évaluer la qualité de la segmentation. Étant donné que l'entropie mesure le degré de désorganisation du contenu de l'image, il est "naturel" qu'elle soit utilisée dans ce contexte. Leur critère s'exprime par la formule suivante :

$$E(I) = H_\ell(I) + H_r(I) \tag{II.26}$$

avec

$$H_\ell(I) = -\sum_{i=1}^{N_R} \frac{|R_i|}{S_I} * log \frac{|R_i|}{S_I}$$

$$H_r(I) = \sum_{i=1}^{N_R} \left(\frac{|R_i|}{S_I} \right) * H(R_i)$$

$$H(R_i) = -\sum_{m \in V_i} \frac{L_i(m)}{|R_i|} * log \frac{L_i(m)}{|R_i|}$$

$H_\ell(I)$ et $H_r(I)$ expriment respectivement l'entropie d'agencement (*layout*) et l'entropie attendue de I alors que $H(R_i)$ donne l'entropie de la région R_i. $L_i(m)$ représente le nombre de pixels de la région R_i ayant pour valeur m.

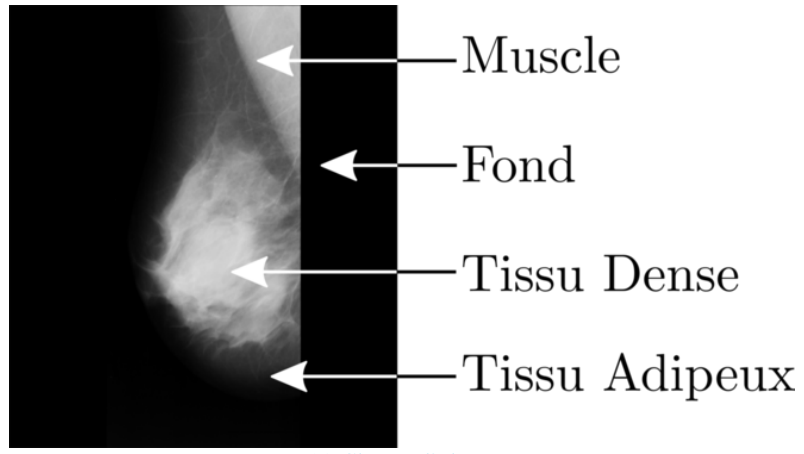
Aujourd'hui plusieurs travaux proposent des bases d'images pour faciliter les validations et universaliser les résultats des travaux en imagerie en général. Le type d'images et les données qui accompagnent les images diffèrent en fonction des caractéristiques des approches à valider. Nous proposons deux types d'images pour valider l'approche proposée.

6.2. Application aux images de mammographies (mini-MIAS)

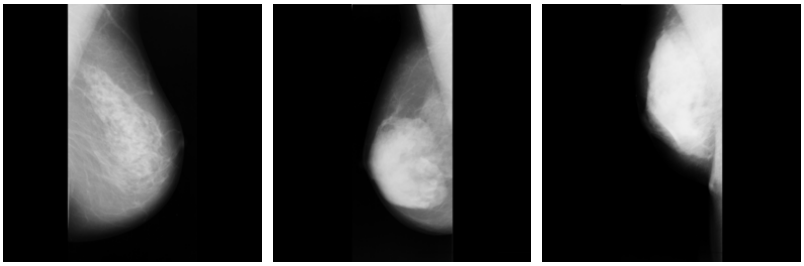
La Mammographic Image Analysis Society (MIAS) est une organisation de groupes de recherche du britannique qui s'intéressent à la compréhension des mammographies et qui a généré une base de données de mammographie numérique. Les clichés sont tirés du programme national de dépistage du cancer du Royaume-Uni et ont été numérisés à `50` pixels de micron avec un microdensitomètre balayage Joyce-Loebl, un dispositif linéaire dans la gamme de densité optique `0-3,2` et représentant chaque pixel avec un mot de 8 bits. La base de données contient 322 films numérisés et disponibles sur `8mm` (Exabyte) bande `2.3GB`. Elle comprend également la "*vérité-terrain*" des radiologues sur les emplacements et le rayon de toutes les

anomalies présentes dans les images. La base de données a été réduite à un bord de pixel de 200 microns et coupée de sorte que toutes les images soient de taille `1024 x 1024` pixels. Les images mammographiques sont disponibles via le **P**ilot **E**uropean **I**mage **P**rocessing **A**rchive (PEIPA) à l'Université d'Essex (Suckling *et al.*, 1994). La figure II.6 présente quelques images exemples de la base mini-MIAS et les objets qui constituent leur contenu sémantique. Le contenu d'une image se divise en quatre types d'objets suivants : le *Fond*, le *Muscle*, le *Tissu dense* et le *Tissu adipeux*.

Des quatre types objets constituant les mammographies, nous avons sélectionné les patchs présentés par la



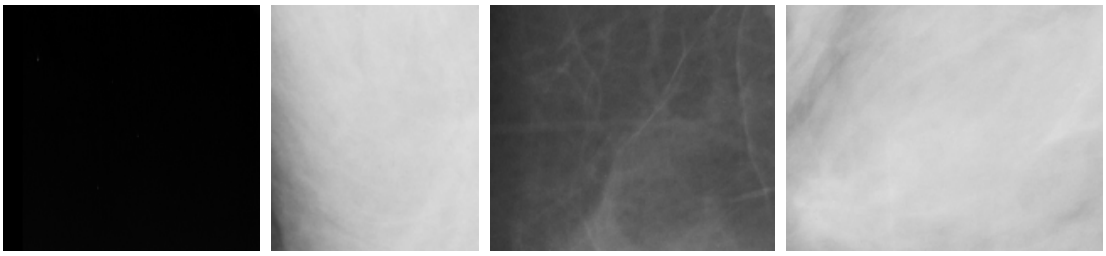
(a) Classes d'objets



(b) Exemple 1 (c) Exemple 2 (d) Exemple 3

FIGURE II.6. – Quelques images de la base mini-MIAS. Une image est typiquement formée du fond, du muscle pectoral, du tissu dense et du tissu adipeux.

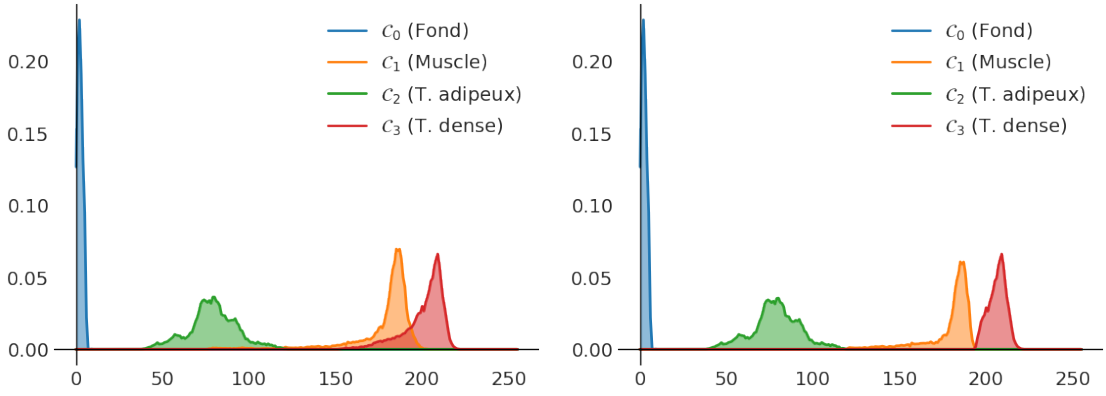
figure II.7.



(a) Patch Fond (b) Patch Muscle (c) Patch T.Adipeux (d) Patch T.Dense

FIGURE II.7. – Schéma illustratif des patchs des classes thématiques.

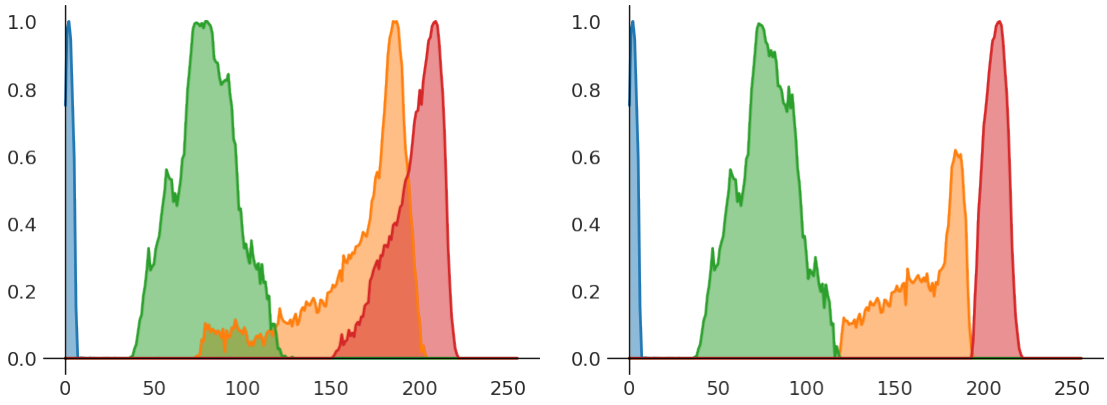
À partir des patchs, nous obtenons les distribution de probabilités correspondantes telles que montrées par la figure II.8.



(a) Distributions de probabilités (b) Distributions de probabilités ajustées

FIGURE II.8. – Schéma illustratif des distributions de probabilités des classes thématiques, initiales et après réajustement.

Afin de pouvoir appliquer le raisonnement possibiliste, les probabilités précédentes sont transformées en possibilités par la transformation de Dubois-Prade(§,5.1). La figure II.9 présente les distributions correspondantes obtenues.



(a) Distributions de possibilités (b) Distributions de possibilités ajustées

FIGURE II.9. – Schéma illustratif des distributions de possibilités des classes thématiques, initiales et après réajustement.

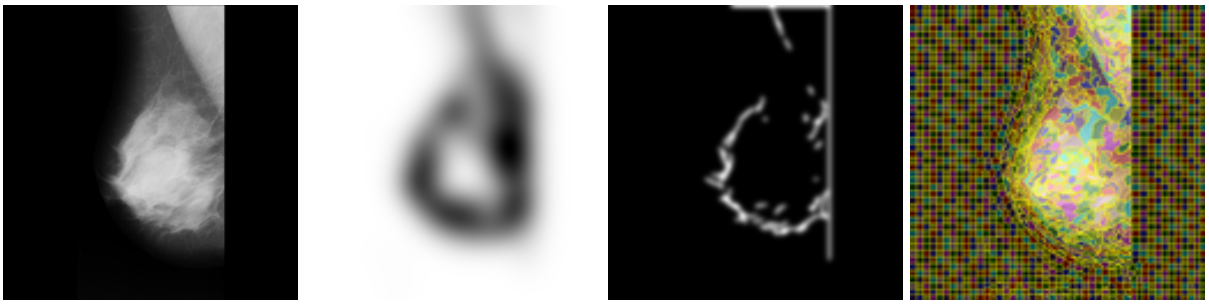
Étant donné les classes d'objets (*Fond*, *Muscle*, *Tissu dense* et *Tissu adipeux*) qui constituent le contenu des image de la base MIAS, la table II.4 présente les configurations spatiales qui sont générées. Sachant que les caractéristiques des contraintes visuelles sont extraites lors du processus de segmentation sur l'image source à segmenter, quelques exemples sur les images de la base MIAS sont présentés par la figure II.10.

Sélection de superpixels germes

La phase de sélection de germes produit les résultats affichés dans la figure II.11. La la première ligne donne les courbes des probabilités et la seconde ceux des possibilités.

TABLE II.4. – Tableau des configurations spatiales possibles des objets dans les images de mammographies.

Superpixels	**Relation spatiale**	
	Voisin	**Inclusion**
Fond	Muscle, Tissu adipeux, Tissu dense.	—
Muscle	Fond, Tissu adipeux, Tissu dense.	—
Tissu adipeux	Fond, Muscle, Tissu dense.	Tissu dense
Tissu dense	Fond, Muscle, Tissu adipeux.	—



(a) Image source (b) Silhouette (c) Contours globaux (d) Superpixels initiaux

FIGURE II.10. – Schéma illustratif des contraintes visuelles pour les images mammographiques et initiale en superpixels décomposition.

Pour constater la qualité de la sélection par classe d'objets dans l'espace probabiliste, l'image de gauche de la figure II.12 montre les distributions des écarts inter-classes utilisés pour la méta-classification. D'autre part la limite des superpixels sélectionnés est tracée par la droite $y = \frac{2}{K}$ (en rouge sur le graphique). La seconde image (celle de droite) montre les courbes représentant les valeurs maximales des superpixels germes. En outre, pour chaque classe, le rectangle de diagramme en boite indique les superpixels sélectionnés pour chaque classe d'objets. Dans cette figure, la largeur du digramme en violon indique la proportion des superpixels ayant la valeur correspondante. Pour la deuxième, les superpixels sélectionnés sont délimités par le rectangle du diagramme en boites.

À partir des analyses probabiliste précédentes, la figure II.13 présente les caractérisations correspondantes dans l'espace de possibilités. Ces courbes montrent bien que l'approche du raisonnement possibiliste offre une meilleure séparation des superpixels. Ainsi, pour les mêmes critères, la distribution de possibilités, bien que présentant les mêmes allures des courbes, fournit une meilleure sélection des germes.

Propagation des connaissances

L'approche de segmentation présentée dans ce chapitre opère par propagation des connaissances des superpixels reconnus vers le reste de de l'image. La figure II.14 présente l'évolution de ce processus sur des images de test.

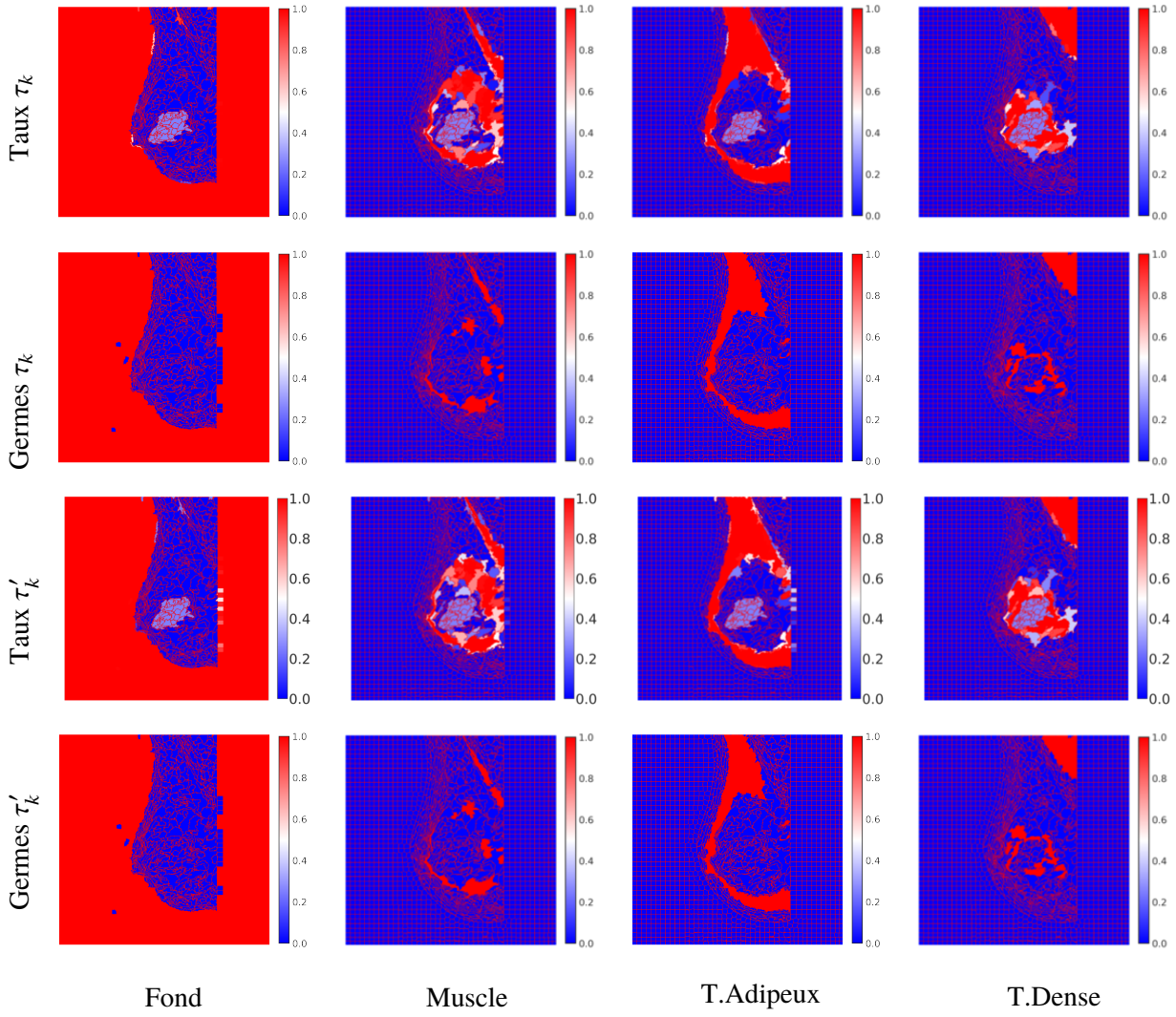


FIGURE II.11. – Schéma du processus de focalisation. Cartes des superpixels germes selon la classe thématique.

Le tableau II.5 présente les résultats de l'évaluation de la segmentation par propagation sur les images selon les critères choisis. Les résultats de la croissance numérique (N) présentent de bonnes performance en termes de Q et Ros. Cela montre l'homogénéité des caractéristiques bas-niveaux utilisées dans cette approche. La segmentation par probabilités Pr et par possibilités π offrent les meilleurs résultats en termes de E et de Ros. En effet, la propagation des connaissances floue (Pr et π), en plus des comparaisons numériques, utilisent une représentation "*plus haut-niveau*" des superpixels.

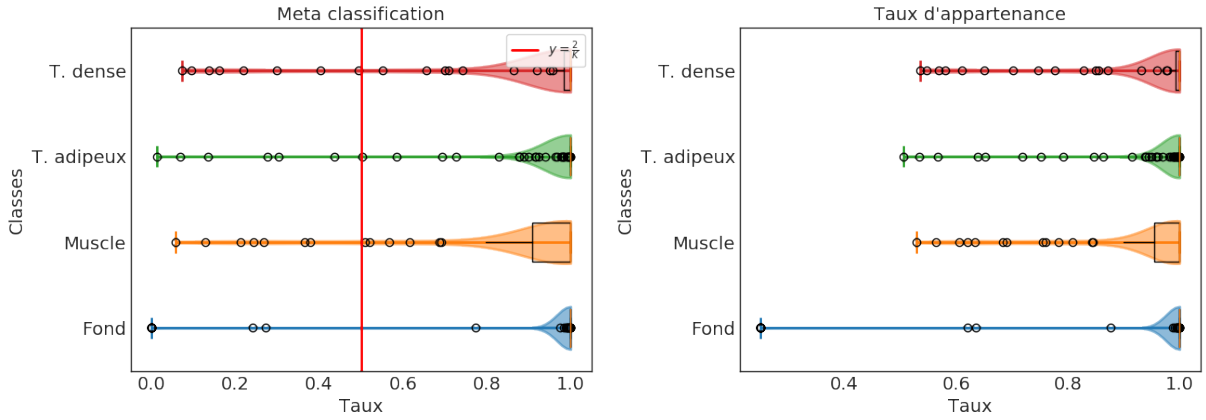


FIGURE II.12. – Séparabilité des germes probabilistes : La figure de gauche donne les écarts de taux d'appartenance pour la méta-classification. Celle de droite indique la distribution des valeurs maximales des classes.

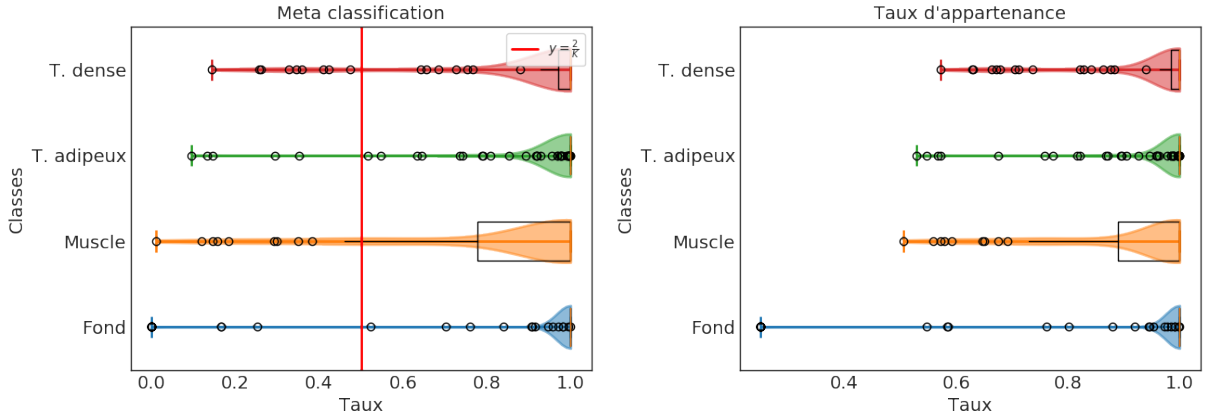


FIGURE II.13. – Séparabilité des germes possibilistes : La figure de gauche donne les écarts de taux d'appartenance pour la méta-classification. Celle de droite indique la distribution des valeurs maximales des classes.

TABLE II.5. – Évaluation des résultats de la segmentation. La première ligne présente les résultats de la croissance numérique. Les résultats de la croissance thématique probabiliste sont donnés dans la deuxième ligne et ceux possibiliste dans la troisième ligne.

		Q Borsotti *et al.* (1998) ↑			*Ros* Rosenberger et Chehdi (2000) ↓			*E* Zhang *et al.* (2003) ↑		
		Min	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**
MIAS	N	0.273	0.865	0.523	0.146	0.215	0.169	0.803	0.876	0.857
	Pr	0.273	0.865	0.523	0.146	0.215	0.169	0.803	0.876	0.857
	π	0.560	0.898	0.783	**0.031**	**0.065**	**0.043**	**0.948**	**0.982**	**0.974**

7. Conclusions

Ce chapitre a présenté notre approche de segmentation par intégration itérative des connaissances spécifiques afin d'orienter progressivement le processus. L'aspect itératif de ce type d'approche leur permet de facilement s'intégrer dans un contexte d'interprétation de scène où elle pourra utiliser les connaissances disponibles et/ou générées par l'étape d'interprétation. Ainsi, l'approche présentée est essentiellement

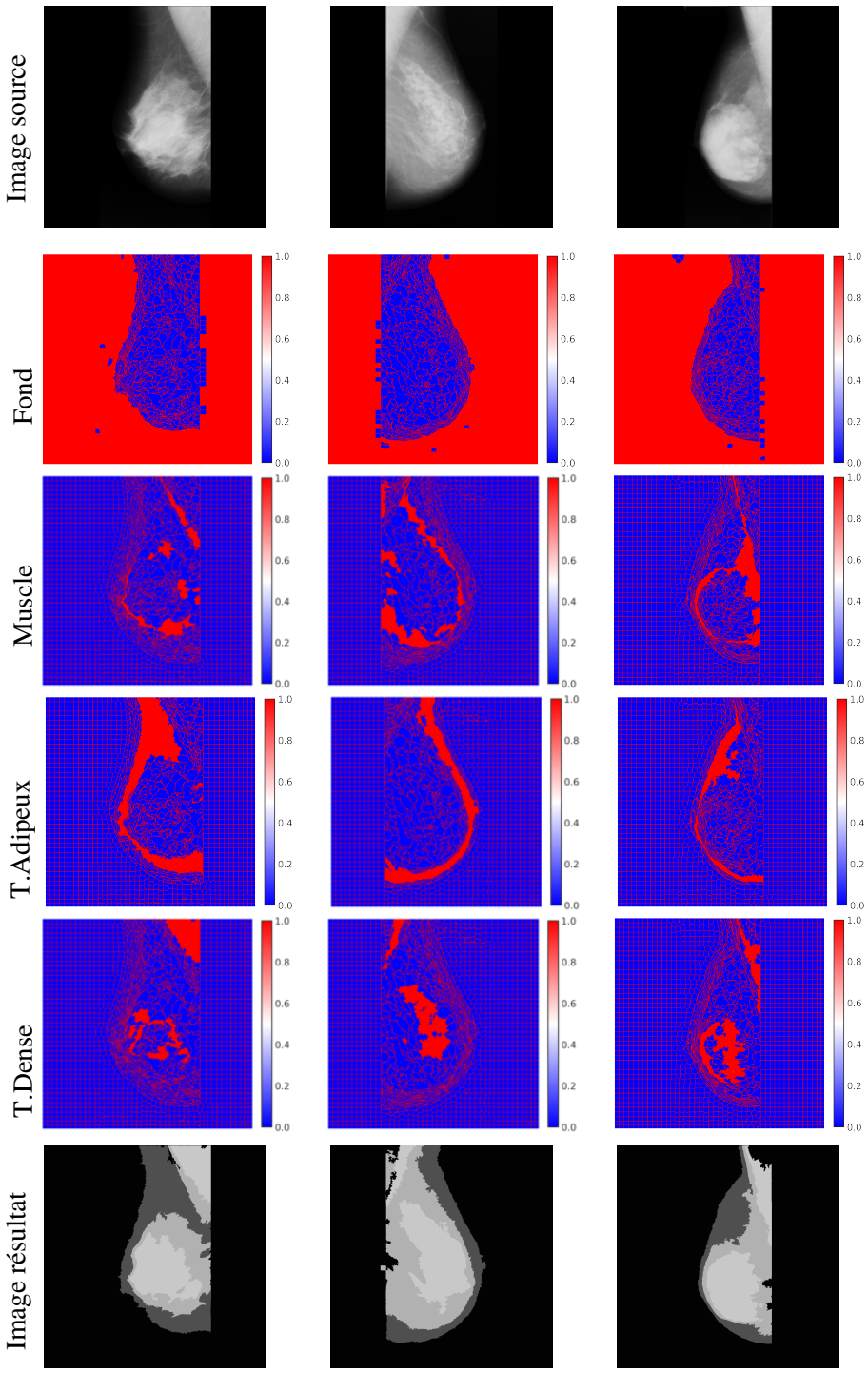


FIGURE II.14. – Résultats du processus de segmentation par propagation. Les premières lignes présentent les cartes des superpixels germes selon la classe thématique. La dernière ligne présente les images segmentées.

composée d'une étape de **Focalisation** d'attention suivie de celle de **Propagation et intégration** de connaissances, et enfin d'une étape de **Fusion** des régions. Ces trois étapes sont successivement répétées jusqu'à stabilisation des régions.

L'étape de focalisation permet de sélectionner, parmi les superpixels initiaux, ceux qui possèdent des

connaissances suffisantes permettant de les caractériser avec la moindre erreur. Ces superpixels sont appelés les germes. À partir des germes, l'étape de propagation utilise les connaissance disponibles au niveau des germes afin de déterminer la nature du reste des superpixels, jusqu'ici non qualifiés. Une étape de fusion permet de regrouper les superpixels assignés aux mêmes classes thématiques pour former les régions de l'image segmentée.
D'abord, nous avons proposé un raisonnement probabiliste qui utilise les atouts de cette théorie, par rapport aux représentation purement numériques, pour aboutir à la segmentation de l'image. Les résultats de l'évaluation permettent de constater le gain obtenu. Ensuite, nous avons remplacé les distribution des probabilités par celles de possibilités afin de gérer l'imprécision de connaissances. Cet ajout a permis de mieux séparer les distributions des classes initiales et ainsi permettre une meilleure classification des superpixels. Les résultats comparatifs présentent une validation de cette hypothèse.

Dans le contexte de segmentation par croissance de régions, la mesure de similarité entre les région est une composante très importante puisqu'elle détermine les régions à fusionner. Cependant au fur des itérations, les régions comparées deviennent plus grandes et plus hétérogènes. La mesure de la similarité perd alors de précision. Le prochain chapitre propose une mesure de similarité multi-échelle et un critère de fusion adaptatif pour tenter de corriger ce problème.

III. Croissance des régions adaptative

Ce chapitre traite de l'aspect itératif de l'approche de segmentation exposée dans le chapitre II. Il soulève deux principales questions : (1) comment calculer la similarité entre les régions pour effectuer une fusion précise et (2) dans quel ordre ces régions doivent être fusionnées. Ainsi, il présente une méthode adaptative de croissance des régions proposant une nouvelle mesure de similarité multi-échelle entre les régions et une stratégie de fusion adaptative pour guider efficacement le processus de fusion des régions. La première composante permet une comparaison de régions à la fois au niveau du contenu et au niveau des frontières communes tandis que la seconde utilise un critère de fusion adaptatif pour garantir que les meilleures agrégations de régions soient faites en premier. Ceci permet de mieux contrôler l'évolution des régions pour offrir des résultats de la segmentation plus cohérents.

Sommaire

1. Introduction

La segmentation d'images est une tâche fondamentale dans de nombreuses applications de reconnaissance de formes et de vision par ordinateur telles que la détection d'objets, la recherche d'images basée sur le contenu et l'analyse d'images médicales. La segmentation est le processus qui consiste à partitionner une image en régions homogènes de pixels avec des caractéristiques similaires et des contours spatialement précis (Haralick et Shapiro, 1985). Malgré la simplicité de sa définition, la segmentation d'image est un problème difficile qui n'a pas de solution universelle. Par ailleurs, la même image peut avoir plusieurs segmentations possibles (Yang *et al.*, 2008) pour au moins deux raisons, selon Tu et Zhu (Tu et Zhu, 2002) : (1) il est fondamentalement complexe de modéliser la grande quantité de motifs visuels d'images, (2) la perception est intrinsèquement ambiguë. En effet, très souvent, on peut fournir différentes interprétations logiques pour une même image.

La croissance de régions est une technique de segmentation populaire basée sur la région qui consiste à fusionner des régions avec les pixels voisins similaires, de manière itérative. A chaque itération, tous les pixels qui bordent la région en croissance sont examinés et les plus similaires sont ajoutés à cette région. Ce processus est illustré par la figure III.1. Les régions initiales peuvent être des pixels ou des régions produites par des techniques de sur-segmentation dédiées, auquel cas elles sont appelées superpixels (Ren et Malik, 2003). Un superpixel est communément défini comme une région atomique perceptuelle obtenue en agrégeant des pixels voisins en fonction de critères de similarité spatiale et d'apparence. Ces dernières années, les techniques de segmentation d'images basées sur les superpixels ont suscité un grand intérêt au sein de la communauté du traitement de l'image, principalement pour leur efficacité de calcul. Les superpixels permettent également une extraction plus efficace des caractéristiques sémantiques contrairement aux patchs d'image (Machairas *et al.*, 2016). Cependant, ces techniques posent deux problèmes principaux : la mesure de la similarité entre les superpixels et les dépendances des superpixels. La mesure de similarité fait référence à la quantification de la similarité entre deux superpixels. Ceci est généralement calculé par une distance normalisée entre les superpixels. Mehnert et Jackway ont déclaré qu'un algorithme de croissance de régions dépend intrinsèquement de l'ordre de traitement des pixels de l'image. Soit à chaque fois que plusieurs pixels sont à la même distance de leurs pixels voisins ou lorsqu'un pixel est à la même distance de plusieurs régions voisines.

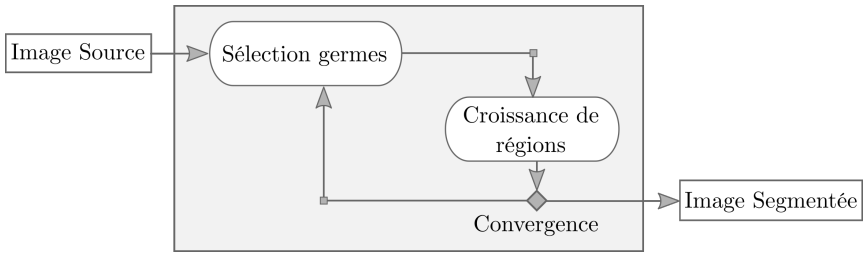


Figure III.1. – Schéma sommaire du processus de la de segmentation par croissance de régions.

Comme solution aux problèmes susmentionnés, ce chapitre propose une mesure de similarité robuste entre les régions ainsi qu'une stratégie de fusion adaptative des régions. La mesure de similarité intègre le contenu et les informations sur les frontières pour fournir une similarité robuste entre les régions. La

stratégie de fusion utilise un critère mis à jour automatiquement pour agréger de manière itérative les régions en fonction de la mesure de similarité proposée.

2. Segmentation par croissance de régions

La segmentation d'images par croissance de régions est une technique largement étudiée dans la littérature. Dans ce chapitre nous abordons précisément deux questions importantes liée à cette façon de segmenter les images à savoir : (i) la mesure de similarité des régions lors du regroupement et (ii) la stratégie de fusion des régions. Ces dernières constituent l'étape de croissance de germes du processus de segmentation par croissance de régions, illustré par la figure III.1.

2.1. État-de-l'art

Plusieurs travaux de recherche se sont évidemment penchés sur ces questions. Ainsi, les auteurs de (Hsu et Ding, 2013) ont proposé une approche de segmentation pour les images naturelles basée sur le **C**ompression-based **T**exture **M**erging (CTM). Ils regroupent d'abord les superpixels par un algorithme de clustering basé sur la compression avec perte, puis effectuent une étape de fusion en utilisant les similarités entre clusters. Ils caractérisent les clusters uniquement par la texture qu'ils extraient à travers des mixtures Gaussiennes. La similarité entre clusters est calculée par une approximation de la distance de Mallows comme suit :

$$d_m(N(\theta_1,\Sigma_1),N(\theta_2,\Sigma_2))^2 = (\theta_1-\theta_2)^T(\theta_1-\theta_2) \tag{III.1}$$

L'utilisation de la compression avec perte leur permet d'introduire la distorsion afin d'obtenir une hiérarchie des segmentations d'une image à plusieurs échelles de quantification. Ils ont également proposé un critère heuristique simple pour déterminer de manière adaptative la distorsion de chaque image et choisir la meilleure segmentation.

Par ailleurs, ces mêmes questions sont abordées dans (Yu *et al.*, 2013) qui propose une segmentation d'images de télédétection se basant sur trois lois de Gestalt définies, pour deux superpixels P et Q, comme suit :

- *le voisinage* qui définit le voisinage spatial de P et Q.
- *la similarité* qui calcule la similarité entre P et Q par : $\exp\left(\left\|\frac{|P|-|Q|}{|P|+|Q|}\right\|_1\right) * \|F(P)-F(Q)\|_1$.
- *la cohérence de couleur* qui caractérise l'homogénéité de la couleur entre P et Q.

$\|\cdot\|_1$ désigne la norme ℓ_1 de l'argument donné. $|P|$ calcule le nombre de pixels dans le superpixel P alors que $F(P)$ calcule son vecteur descripteur. Ce dernier comprend plusieurs composante notamment la luminosité, la texture, les contours, les informations spatiales et le contexte local. Les auteurs de (Yu *et al.*, 2013) proposent de plus un algorithme de regroupement de superpixels à deux niveaux : fusion grossière et fusion fine. Le premier niveau se concentre sur l'accélération de l'exécution de l'algorithme de segmentation, tandis que le second s'occupe de l'amélioration de la précision de la segmentation.

Dans le même ordre d'idées, une approche de détection d'action dans une séquence vidéo est proposée dans (Oneata *et al.*, 2014). A partir des superpixels initiaux représentés à l'aide d'un graphe, les auteurs proposent un algorithme d'agrégation hiérarchique, spatiale et temporelle. La similarité S_{PQ} entre deux

clusters de superpixels, P et Q, est exprimée en fonction du poids ω de l'arête reliant P et Q.

$$S_{PQ} = \frac{1}{|\mathcal{B}(P,Q)|} * \sum_{p,q \in \mathcal{B}(P,Q)} \omega(p,q) \tag{III.2}$$

où $\mathcal{B}(P,Q)$ dénote l'ensemble des arêtes reliant P et Q. Le poids ω est formé de trois composantes qui expriment d'abord le poids spatial direct entre P et Q par ω^{sp}, ensuite le poids spatial du second ordre par ω^{2hop} et enfin le poids temporel par ω^{t}. Cette dernière composante est introduite pour capturer l'aspect temporel entre les frames de la séquence vidéo.

$$\begin{aligned} \omega^{sp}(P,Q) &= \alpha_{\mu} d_{\mu}(P,Q) + \alpha_{col} d_{col}(P,Q) \\ &\quad + \alpha_{flow} d_{flow}(P,Q) + \alpha_{mb} d_{mb}(P,Q) \\ &\quad + \alpha_{edge} d_{edge}(P,Q) \end{aligned} \tag{III.3}$$

$$\begin{aligned} \omega^{2hop}(P,Q) &= \alpha_{\mu} d_{\mu}(P,Q) + \alpha_{col} d_{col}(P,Q) \\ &\quad + \alpha_{flow} d_{flow}(P,Q) + \alpha_{2hop} \end{aligned} \tag{III.4}$$

$$\begin{aligned} \omega^{t}(P,Q) &= \alpha^{t}_{\mu} d_{\mu}(P,Q) + \alpha^{t}_{col} d_{col}(P,Q) \\ &\quad + \alpha^{t}_{flow} d_{flow}(P,Q) \end{aligned} \tag{III.5}$$

où $d_{\mu}(P,Q) = min(|\mu(P) - \mu(Q)|, 30)$ correspond à la distance seuillée entre les moyennes des couleur P et Q. $d_{col}(P,Q)$ et $d_{flow}(P,Q)$ sont respectivement les distances entre les histogrammes de couleur et de flow calculées par la formule du chi-carré (χ^2). $d_{mb}(P,Q)$ et $d_{edge}(P,Q)$ sont les distances géodésiques entre les centroïdes des deux clusters de superpixels. La formule III.2 est utilisée pour apprendre un classificateur qui guide le processus de fusion aléatoire pour aboutir à la segmentation des frames.

Plus récemment, dans (Yang *et al.*, 2016) une approche de segmentation d'image basée sur le regroupement est proposée. Dans ce travail, Yang *et al.* caractérisent les superpixels par un vecteur de `74` descripteurs dont `64` pour la couleur et `10` pour la texture. En particulier, une nouvelle mesure de similarité basée sur un noyau Gaussien flou est proposée. En effet, pour deux superpixels P et Q de vecteurs de degrés d'appartenance aux clusters respectifs u et v, la similarité S_{PQ} est calculée comme suit :

$$S_{PQ} = \begin{cases} 0, & \texttt{si } P \texttt{ et } Q \texttt{ ne sont pas t-plus-proches-voisins l'un de l'autre.} \\ 1, & \texttt{si } P \texttt{ et } Q \texttt{ appartiennent au même cluster et sont } t\texttt{-plus-proches-voisins l'un de l'autre.} \\ e^{[Ln(2) \times (u \odot v)]} - 1, & \texttt{sinon} \end{cases} \tag{III.6}$$

avec $\odot$ l'opérateur du produit entre deux vecteurs de réels. Cette formulation de la similarité est introduite pour réduire la perte de sémantique dans la représentation par le noyau Gaussien lorsque la taille du vecteur de caractéristiques devient grande.

Il faut noter que, dans tous ces travaux, le regroupement des superpixels nécessite la validation de deux critères clés : la connexité spatiale et la ressemblance visuelle. Cela sous-entend qu'une approche efficace permettrait de choisir le voisin spatial le plus similaire d'un superpixel de manière locale mais aussi globale. Ainsi, ce chapitre propose une mesure de similarité robuste entre les superpixels qui combinent efficacement le contenu et les frontières des superpixels pour fournir une mesure précise. De plus, nous définissons une stratégie de fusion pour guider le processus d'agrégation des superpixels en introduisant un critère de fusion adaptatif et un ordre de priorité entre les regroupements de superpixels.

2.2. Le modèle proposé

Premièrement, nous proposons en tant qu'initialisation une décomposition de l'image source en superpixels en utilisant une extension de l'algorithme **S**imple **L**inear **I**terative **C**lustering (SLIC) (Achanta *et al.*, 2012). SLIC est populaire car il permet de générer des superpixels avec une implémentation simple, une exécution rapide et une assez bonne précision. Cependant, on peut observer qu'il échoue parfois à respecter parfaitement les contours de l'image. Par conséquent, des contraintes de contour globaux sont appliquées (§ 3) sur l'algorithme SLIC pour résoudre ce problème. Une fois qu'une décomposition de superpixel précise est obtenue, la croissance itérative de superpixels commence. En particulier, les superpixels sont regroupés en régions [1] selon deux critères principaux : la sélection mutuelle des superpixels et le chevauchement des contours globaux. Chaque région choisit le meilleur superpixel candidat pour la fusion à partir de ses voisins en utilisant la mesure de similarité (§ 5.1), qui est basée sur les caractéristiques de superpixel appropriées. Par la suite, chaque couple d'une région et son superpixel voisin qui se sont mutuellement choisis avec suffisamment de similarité et qui ne sont pas séparés par un contour global sont regroupés. Ce processus de fusion itératif est répété jusqu'à ce que le seuil de similarité final soit atteint.

Noter qu'à la première boucle, chaque superpixel est considéré comme une région candidate fusionnante. La sélection mutuelle et les critères de contour globaux garantissent que chaque région soit fusionnée avec le meilleur superpixel de son voisinage à une itération donnée.

L'algorithme III.1 décrit l'approche de segmentation proposée et la figure III.2 illustre sommairement son principe.

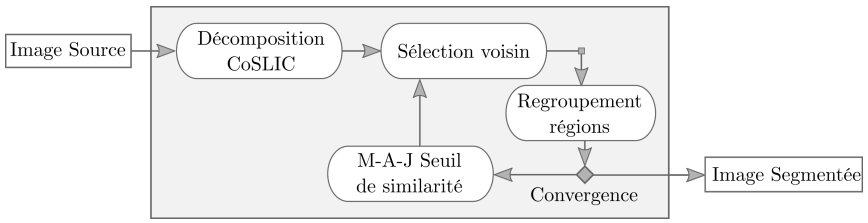


FIGURE III.2. – Schéma global de l'approche de segmentation par croissance de régions adaptative.

3. CoSlic : extension de SLIC par contours globaux

Malgré son efficacité, l'algorithme SLIC (Achanta *et al.*, 2012) produit souvent des superpixels qui chevauchent plusieurs régions sur les contours de l'image d'entrée. Dans le contexte des approches

1. Nous appelons région un ensemble de superpixels regroupés en zone homogène. A noter qu'à l'initialisation du processus de regroupement les régions sont formées chacune d'un seul superpixel. Ainsi ces deux termes sont parfois interchangeables dans ce manuscrit.

Algorithme III.1 : Segmentation par croissance de régions adaptative

Données : $\mathcal{I}$,$\mathcal{S}_0$; ▷ $\mathcal{I}$:Image source,$\mathcal{S}_0$:Similarité d'arrêt
Résultat : $\mathcal{S}_\mathcal{P}$; ▷ $\mathcal{S}_\mathcal{P}$:ensemble de superpixels

début
- $\mathcal{G}_C := \text{Canny}(\mathcal{I})$;
- $\mathcal{S}_\mathcal{P} := \text{CoSLIC}(\mathcal{I})$;
- $\mathcal{S}_{it} := 1$; ▷ $\mathcal{S}_{it}$:Similarité adaptative
- $Ms := true$;
- **tant que** $Ms = true \quad and \quad \mathcal{S}_{it} \geq \mathcal{S}_0$ **faire**
  - $Ms := false$;
  - $Mc := \emptyset$;
  - **pour** $P \in \mathcal{S}_\mathcal{P}$ **faire**
    - $N := argmax_i\,(\text{Sim}(P, N_i))$; ▷ $\mathcal{N}(P)$:voisins de P
      $\forall \quad N_i \in \mathcal{N}(P) \quad and \quad \text{Sim}(P, N_i) \geq \mathcal{S}_{it}$;
    - $Mc := Mc \cup (P, N)$;
  - **pour** $(P, Q) \in Mc$ **faire**
    - **si** $(Mc(P) = Q \quad and \quad Mc(Q) = P)$ **alors**
      - $Ms := true$;
      - **si** $((P \cup Q) \cap Gc = \emptyset)$ **alors**
        - $\mathcal{S}_\mathcal{P} := \text{Merge}(\mathcal{S}_\mathcal{P}, P, Q)$;
        - $\mathcal{S}_{it} := \alpha_{it} * \mathcal{S}_{it}$; ▷ α_{it}:coefficient de m-a-j(§5.4)
      - **sinon**
        - $\mathcal{N}(P) := \mathcal{N}(P) - Q$;
        - $\mathcal{N}(Q) := \mathcal{N}(Q) - P$;
        - $\mathcal{S}_{it} := \frac{1}{\alpha_{it}} * \mathcal{S}_{it}$;

de segmentation basées sur les superpixels, ce défaut peut conduire à de très mauvais résultats de la segmentation. Pour corriger cette faiblesse nous proposons une extension de cet algorithme en tirant parti des détecteurs de contours classiques pour pallier cet inconvénient. Pour ce faire, à partir de l'image source, la carte de contours globaux est extraite à l'aide du détecteur de contours de Canny (Canny, 1986). Ensuite, pour chaque superpixel, nous effectuons une vérification croisée avec les contours globaux. Les superpixels traversés par des contours globaux sont divisés le long des contours en superpixels plus petits.

Ainsi, étant donné une image d'entrée $\mathcal{I}$, nous notons $\mathcal{S}_p$ le superpixel généré par SLIC sur $\mathcal{I}$, et $\mathcal{G}_C$ la carte de contours globaux obtenue en appliquant le détecteur de contours de Canny sur $\mathcal{I}$. Notre algorithme de décomposition en superpixels CoSLIC partitionne chaque superpixel $P_i \in \mathcal{S}_p$ en sous-superpixels P_{i_j}

basé sur $\mathcal{G}_C$ tel que les trois conditions suivantes soient remplies :

$$\begin{cases} (a) & P_{i_j} \cap \mathcal{G}_C = \emptyset, \forall i, j \\ (b) & \bigcup_j P_{i_j} = P_i \\ (c) & P_{i_j} \cap P_{i_k} = \emptyset, \forall j \neq k \end{cases} \tag{III.7}$$

La première condition garantit qu'aucun superpixel ne soit traversé par $\mathcal{G}_C$. Les deux dernières conditions définissent les sous-superpixels P_{i_j} obtenus en tant que partition du superpixel parent P_i. La figure III.3 montre un schéma illustratif de l'extension proposée pour le SLIC : CoSLIC.

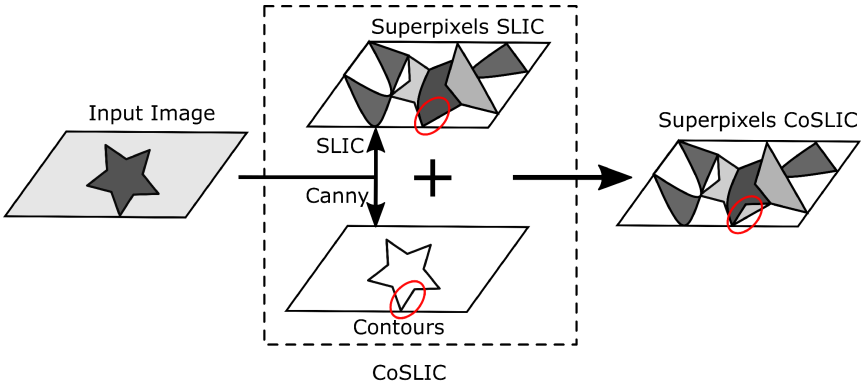


FIGURE III.3. – Schéma illustratif de CoSLIC : une extension de SLIC (Achanta *et al.*, 2012) en utilisant les contours globaux pour produire une meilleure adhérence aux contours.

4. Descripteurs multi-niveaux de superpixel

Dans un contexte de segmentation par regroupements progressifs, la taille des régions regroupées varie avec les itérations. Ce changement implique un déséquilibre de caractérisation selon la taille des régions. Ainsi, les caractéristiques extraites pour les régions de petite taille seront assez précis tandis que celle de grande taille seront moins précis. Afin de réduire ce déséquilibre, nous proposons de calculer les caractéristiques des régions à différentes échelles de leur taille.

Le processus de croissance de superpixels vers une segmentation précise peut être représenté par une structure arborescente où les nuds correspondent à des régions formées par des groupes de superpixels. Cet arbre hiérarchique est construit de manière ascendante dans lequel deux régions voisines similaires sont regroupées en un nud de région parent. Étant donné l'image d'entrée $\mathcal{I}$, initialement partitionnée en un ensemble de superpixels $\mathcal{S}_\mathcal{P}$, cette définition arborescente de l'approche de croissance de superpixels peut être exprimée comme une succession de partitions ordonnées Γ_k de $\mathcal{I}$ à différents niveaux k, allant de 0 à K. La racine de l'arbre est notée par le plus haut niveau K et les feuilles, correspondant aux superpixels initiaux, sont définies par le niveau 0. Ainsi, nous avons

$$\forall \quad i, j \in [0, K], \quad \begin{cases} i \neq j \implies \Gamma_i \neq \Gamma_j \\ i < j \implies \|\mathcal{R}(\Gamma_i)\| \geq \|\mathcal{R}(\Gamma_j)\| \end{cases} \tag{III.8}$$

où $\mathcal{R}(\cdot)$ définit l'ensemble des régions formées par les agrégations de superpixels et $\|\cdot\|$ indique le nombre d'éléments dans un ensemble.

Les superpixels sont le seul niveau à être directement lié à l'image dans ce processus hiérarchique. Pour cette raison, les superpixels sont décrits par des caractéristiques extraites directement de l'image. L'image est représentée dans l'espace colorimétrique CIE Lab. La littérature (Tkalcic et Tasic, 2003) a établi cet espace de couleur comme le mieux corrélé à la perception visuelle humaine. De plus, les distances vectorielles classiques telles que les distances euclidiennes se révèlent visuellement significatives dans l'espace colorimétrique CIE Lab. Un superpixel P est décrit par un vecteur de dix caractéristiques $F_P = (f_p^0, \ldots, f_p^9)$ divisé en trois catégories de descripteurs : intensité, texture et gradient. Les caractéristiques d'intensité comprennent la moyenne, la variance, l'asymétrie et un histogramme des valeurs d'intensité sur chacune des composants L, a et b. Le contraste, la corrélation, l'énergie et l'entropie sont extraits en tant que caractéristiques de texture tandis que l'histogramme des orientations et l'histogramme des amplitudes sont calculés à partir du gradient de l'image tel que résumé dans le tableau III.1.

TABLE III.1. – Caractéristiques calculées pour chaque superpixel. Les caractéristiques sont divisées en descripteurs d'intensité, de texture et de gradient.

Catégorie	**Caractéristiques**	
Intensité	Moyenne	f^0
	Variance	f^1
	Skewness	f^2
	Histogramme	f^3
Texture	Contraste	f^4
	Corrélation	f^5
	Énergie	f^6
	Entropie	f^7
Gradient	Histogramme d'orientations	f^8
	Histogramme de magnitude	f^9

Les régions sont caractérisées par une concaténation de leurs régions descendantes directes dans le modèle hiérarchique selon leur ordre d'apparition. Ceci décrit la séquence des régions filles qui sont utilisées pour former la région et donne une représentation multi-échelle de la région considérée. Cette idée a été initialement utilisée pour représenter un nud feuille, dans une structure arborescente, par une séquence de ses ascendants dans (Conze *et al.*, 2017). Dans l'approche proposée, cette idée est appliquée de manière ascendante pour s'adapter à notre approche et garder le même effet. Ainsi, chaque région a une représentation à plusieurs niveaux qui contient la description de tous ses descendants ordonnés selon leur niveau dans sa hiérarchie. Dans cet espace de description, la région BC dans la figure III.4-b sera caractérisée par $F_{BC} = (f_B, f_C)$ où $F_B = (f_B^0, \ldots, f_B^9)$ et $F_C = (f_C^0, \ldots, f_C^9)$ décrivant respectivement les superpixels B et C.

5. Croissance de régions adaptative

5.1. Similarité des régions

L'efficacité de la mesure de similarité entre régions dans une approche de segmentation par croissance de région est un composante cruciale. Par ailleurs, au fur et à mesure que la taille des régions augmente leur contenu devient plus hétérogène. Ce changement rend la mesure de moins en moins précise car elle ne s'adapte pas au fur des itérations. Nous proposons dans cette section une mesure qui tient compte de ce changement de taille des régions.
La mesure de similarité assure le regroupement progressif des éléments afin d'obtenir les régions finales de la segmentation. De manière générale, la similarité $S(X,Y)$ entre deux éléments X et Y exprime la ressemblance entre A et B et doit vérifier les trois contraintes suivantes :

— Non négativité : $S(X,Y) \geqslant 0$

— Symétrie : $S(X,Y) = S(Y,X)$

— Autosimilarité maximale : $S(X,X) = 1$

Dans le contexte de segmentation, en général les éléments correspondent aux régions à regrouper et sont représentées par leur caractéristiques extraites. Deux grandes familles de similarité regroupent les diverses formules qui ont été proposées dans la littérature pour capter la similarité entre deux régions (Bouchon-Meunier *et al.*, 2008 ; Gan *et al.*, 2007) :

— les mesures *géométriques*. Elles considèrent les informations à comparer comme étant des points dans un espace métrique de représentation.

— les mesures *ensemblistes*. Elles considèrent chaque information représentant un objet, comme étant un sous-ensemble algébrique de l'univers des primitives Ω, comportant les primitives réalisées par cette information.

Dans notre approche, une région $R \in \Psi_r$ est définie, à partir d'un ensemble de superpixels $\mathcal{S}_p$, comme un groupement non vide de superpixels. L'équation III.9 définit l'ensemble des régions, Ψ_r.

$$\Psi_r = \{x \mid x \subseteq \{\mathcal{S}_p \setminus \emptyset\}\} \tag{III.9}$$

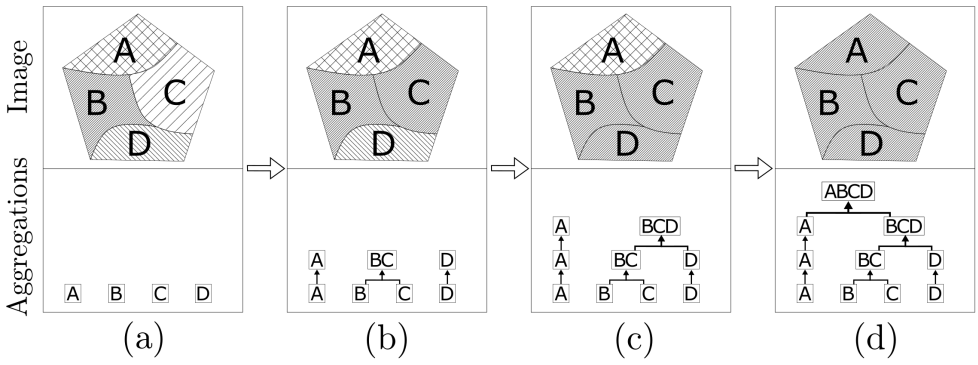


FIGURE III.4. – Processus de croissance hiérarchique de superpixels utilisant une image sur-segmentée en superpixels A, B, C et D. Les agrégations itératives de superpixels sont modélisées à travers une structure hiérarchique ; Les caractéristiques sont concaténées le long d'une telle structure pour atteindre une description robuste de la région à plusieurs échelles.

La mesure de similarité proposée appartient à la famille des mesures géométriques. Elle calcule la différence à la fois au niveau de la région et de la frontière, pour donner une vision globale et locale de la comparaison. L'idée est que deux régions à fusionner doivent avoir un contenu similaire et une frontière commune harmonieuse, ce qui signifie qu'il n'y a pas de changement brusque d'une région à l'autre. Par conséquent, la similarité entre les régions est définie à partir de leur contenu mais aussi de leur frontière commune. La similarité de contenu est calculée comme une comparaison des deux régions, tandis que la similarité de frontière est définie sur la base de la similarité entre les superpixels connectés qui forment la frontière de chaque région. Ainsi, la forme générale de la mesure de similarité entre deux régions, $R_i, R_j \in \Psi_r$, peut être exprimée par l'équation III.10.

$$Sim(R_i, R_j) = F(Sim_C(R_i, R_j), Sim_B(R_i, R_j)) \tag{III.10}$$

où $Sim_C(R_i, R_j)$ et $Sim_B(R_i, R_j)$ dénotent respectivement la similarité du contenu et de frontière entre R_i et R_j.

5.1.1. Similarité du contenu

Tout d'abord, la similarité du contenu entre R_i et R_j, représentés par leurs vecteurs caractéristiques multi-échelles respectives f_i et f_j est définie. En ne considérant que le descripteur ℓ^{th}, une similarité de contenu intermédiaire entre R_i et R_j est également définie comme :

$$Sim_C^{\ell}(R_i, R_j) = \mu(\chi^2(f_i^{\ell}, f_j^{\ell})), \tag{III.11}$$

où

- $\mu(x) = \exp\left(\frac{1}{2} * \left(\frac{x}{\sigma}\right)^2\right)$ est l'appartenance floue gaussienne centrée sur zéro. Cette fonction est utilisée ici pour normaliser la valeur de similarité et σ sa valeur d'écart-type.
- La fonction χ^2 provient du test-statistique χ^2 (Snedecor et Cochran, 1967) où elle est utilisée pour tester l'ajustement entre une distribution et les fréquences observées. Elle est définie, pour deux vecteurs $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, comme $\chi^2(x, y) = \sum_{i=1}^{n} \left(\frac{(x_i - y_i)^2}{(x_i + y_i)}\right)$.

En considérant tous les descripteurs, le vecteur des similarités de niveau intermédiaire peut être exprimé comme suit :

$$S^{\ell} = \begin{bmatrix} Sim_C^{\ell}(f_i^0, f_j^0)) \\ \vdots \\ Sim_C^{\ell}(f_i^9, f_j^9)) \end{bmatrix} = \begin{bmatrix} \mu(\chi^2(f_i^0, f_j^0)) \\ \vdots \\ \mu(\chi^2(f_i^9, f_j^9)) \end{bmatrix} = \begin{bmatrix} S_0 \\ \vdots \\ S_9 \end{bmatrix}, S_i \in [0,1], i = (0, \ldots, 9) \tag{III.12}$$

La similarité du contenu finale entre R_i et R_j, $Sim_C(R_i, R_j)$, est définie comme une moyenne de quelques valeurs statistiques de la distribution des similarités intermédiaires entre les deux régions en utilisant tous les descripteurs. Elle est définie par :

$$Sim_C(R_i, R_j) = \frac{1}{4}(S_{max} + S_{min} + S_{mean} + S_{var}) \tag{III.13}$$

avec :

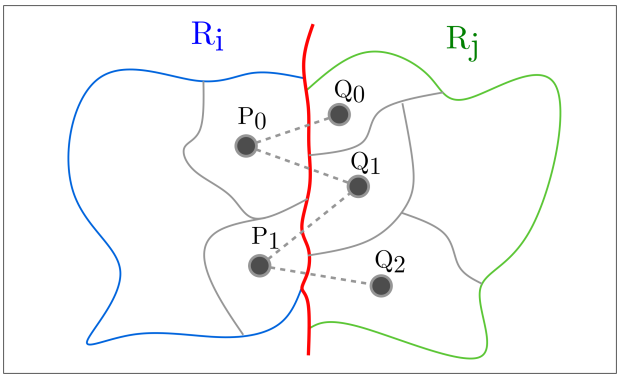


FIGURE III.5. – les deux régions R_i et R_j sont séparées par 4 couples de superpixels voisins. La similarité de frontière est calculée comme $Sim_B(R_i, R_j) = \frac{1}{4}\left[Sim_C(P_0, Q_0) + Sim_C(P_0, Q_1) + Sim_C(P_1, Q_1) + Sim_C(P_1, Q_2)\right]$

- $S_{max} = max_\ell(S^\ell)$, $S_{min} = min_\ell(S^\ell)$ sont les extrema des similarités au niveau des descripteurs entre les deux régions et
- $S_{mean} = mean_\ell(S^\ell)$, $S_{var} = var_\ell(S^\ell)$ fournit la mesure de la tendance centrale et la dispersion de ces similarités en tant que moyenne et variance, respectivement.

Une région est décrite par une gamme de caractéristiques avec des dynamiques différentes. Par exemple, une même valeur de similarité de $S = 2,9$ obtenue séparément pour la texture et pour l'intensité peut ne pas exprimer le même degré de ressemblance au niveau des régions considérées. Ainsi, en normalisant ces similarités intermédiaires au niveau des descripteurs, nous nous assurons que la similarité calculée ne sera pas biaisée par la dynamique des descripteurs. De plus, nous observons expérimentalement que de telles statistiques donnent une meilleure similarité qu'une distance classique, comme la distance euclidienne, entre les similarités intermédiaires.

5.1.2. Similarité de frontière

La similarité de frontière est conçue pour empêcher la fusion des régions adjacentes similaires mais appartenant à différents objets. Deux régions voisines R_i et R_j sont supposées être séparées par une frontière composée de N couples de superpixels que nous désignerons par P_i pour les superpixels qui appartiennent à R_i et Q_j pour ceux de R_j. Ainsi, l'équation III.14 donne l'expression de la mesure de similarité de frontière entre R_i et R_j. Un exemple illustratif est montré dans la figure III.5.

$$Sim_B(R_i, R_j) = \frac{1}{N}\left(\sum_{P_i \in \mathcal{N}(Q_j)}^{N} Sim_C(P_i, Q_j)\right) \tag{III.14}$$

En d'autres termes, $Sim_B(R_i, R_j)$ est considérée comme la moyenne des similarités entre chaque couple de superpixels le long de la frontière séparant R_i et R_j.

5.1.3. Similarité finale

Enfin, la similarité entre les deux régions, $Sim(R_i, R_j)$ exprimée par l'équation III.17, est définie comme une combinaison pondérée des similarités du contenu et de celle de frontière. Le coefficient de pondération

du contenu, défini par :

$$\omega_C = \sqrt{\frac{min\left(|R_i|, |R_j|\right)}{max\left(|R_i|, |R_j|\right)}} \tag{III.15}$$

Il est ajouté afin d'équilibrer la similarité des régions en conséquence par rapport à leur taille respective. Aussi, étant donné C_i (resp. C_j) comme la circonférence de R_i (resp. R_j) et β la longueur de la frontière commune, en nombre de pixels, de R_i et R_j, nous avons défini le poids de $Sim_B(R_i, R_j)$ comme suit :

$$\omega_B = \sqrt{\frac{\beta * (C_i + C_j)}{2 * C_i * C_j}} \tag{III.16}$$

Cette valeur correspond au pourcentage de la frontière commune entre les deux régions considérées.

$$Sim(R_i, R_j) = \omega_C * Sim_C(R_i, R_j) + \omega_B * Sim_B(R_i, R_j) \tag{III.17}$$

Le coefficient de contenu favorise le regroupement de régions de tailles similaires, tandis que le coefficient de bordure favorise le regroupement de régions qui partagent la plus grande frontière. Cela agit comme un régulateur et empêche une catégorie de régions de croître plus vite que d'autres.

5.2. Stratégie de fusion

Dans un processus de segmentation itératif par regroupement de régions, la mesure de similarité et le critère d'homogénéité permettent d'estimer la ressemblance visuelle de régions. Le choix des régions à fusionner est assuré par la stratégie de fusion implémentée. Cette étape permet particulièrement, pour une région donnée, de sélectionner la meilleure région voisine parmi celles qui vérifient le prédicat du critère d'homogénéité. Différentes heuristiques pour la sélection d'une région adjacente ont été proposées ; (Baatz et Shäpe, 2000) en présente une synthèse intéressante.

Étant donné une région P et la liste de ses régions adjacentes, la stratégie de *Fitting (F)* consiste à choisir la première région adjacente Q pour laquelle le prédicat d'homogénéité $\hat{\mathcal{H}}$ est vérifié avec P. Une amélioration de cette première stratégie consiste à choisir une région adjacente Q pour laquelle le coût de fusion avec P est le minimum parmi ceux qui vérifient le prédicat. Cette heuristique est appelée *Best Fitting (BF)*. En considérant, BF et étant donné trois régions adjacentes, P, Q et R. Lorsque P choisit Q comme région de fusion et Q choisit R qui elle même choisit P, on se retrouve dans un cycle fermé infini. Ces régions ne vont jamais être regroupées dans ces conditions. L'heuristique *Local Mutual Best Fitting (LMBF)* ajoute la contrainte supplémentaire de sélection mutuelle. Cette contrainte stipule que deux régions P et Q sont regroupées si et seulement si elles vérifient le prédicat d'homogénéité et que P choisisse Q et Q choisisse P (Lassalle *et al.*, 2015). Pour pallier au problème de conflit entre les régions choisies, la dernière heuristique, *Global Mutual Best Fitting (GMBF)*, impose de ne fusionner qu'une seule paire de régions adjacentes à chaque itération du regroupement. La paire de régions qui fusionne est celle pour laquelle la mesure de similarité est la maximum parmi celles de toutes les autres paires de régions de l'image. Cette heuristique est la plus contraignante car seulement une paire fusionne à chaque itération, ce qui peut augmenter sévèrement le temps d'exécution de l'algorithme (Blaschke et Hay, 2001 ; Lassalle *et al.*, 2015).

Nous proposons, dans ce travail, d'utiliser cette dernière heuristique en ajoutant un ordre de priorité aux fusions itératives : cela conduit à des fusions par agrégations adaptatives.

5.3. Agrégations adaptatives

La procédure de fusion proposée est inspirée de l'algorithme de regroupement agglomératif (Ward Jr, 1963) pour sa simplicité conceptuelle et sa flexibilité. Néanmoins, les algorithmes de croissance de région dépendent intrinsèquement de l'ordre de traitement des pixels de l'image (Mehnert et Jackway, 1997). L'ordre de fusion affecte fortement la convergence de l'approche car si deux régions sont fusionnées à tort, cette erreur sera propagée aux étapes suivantes. Pour résoudre ce problème, nous proposons une stratégie de fusion qui minimise les erreurs à chaque étape. En fait, à une itération donnée, chaque région fait un choix unique de fusion parmi les régions qui lui sont adjacentes. Cette sélection assure un choix optimal local. Une fois que chaque région a fait son choix, une étape de validation effectue une vérification d'optimisation globale pour supprimer toutes les régions qui ne se sont pas mutuellement choisies ou séparées par un contour global. Ainsi, nous évitons tout conflit et garantissons une cohérence maximale à l'étape de regroupement. L'algorithme III.2 résume les principales étapes de la procédure de fusion proposée.

Algorithme III.2 : Agrégations itératives

```
/* Meilleure configuration locale                                   */
pour P ∈ S_P faire
    N := argmax_i (Sim(P, N_i))   ∀ N_i ∈ N(P) and Sim(P, N_i) ≥ S_it ;
    N*(P) := N ;                          ▷ N*(P): meilleur voisin de P
/* Meilleure configuration globale                                  */
pour P ∈ N*  |  N*(P) = Q faire
    si (N*(P) = Q and N*(Q) = P) alors
        si ((P ∪ Q) ∩ GlobalContours = ∅) alors
            A*(P) = Q ;                   ▷ A*(P): agrégation du voisin P
            A*(Q) = P ;
            S_it := α_it * S_it ;
        sinon
            N(P) := N(P) − Q ;
            N(Q) := N(Q) − P ;
            S_it := 1/α_it * S_it ;
/* Ordre de priorité d'agrégations                                  */
pour P ∈ A*  |  A*(P) = Q faire
    si Sim(P, Q) > S_it alors
        S_P := Merge(S_P, P, Q) ;
```

En résumé, la procédure de fusion consiste en trois étapes consécutives. Étant donné une région R_i

1. nous choisissons d'abord la meilleure configuration de fusion en identifiant sa meilleure région

voisine R_j, en utilisant la mesure suivante :

$$\text{Sim}(R_i, R_j) = max\,(\text{Sim}(R_i, N_i)) \quad \forall \quad N_i \in \mathcal{N}(R_i) \tag{III.18}$$

2. Deuxièmement, la meilleure configuration globale de fusion d'image est faite à partir de la meilleure configuration de toutes les paires de régions candidates à la fusion. À cette fin, nous supposons que

$$\mathcal{N}^*(R_i) = R_j \quad \text{et} \quad \mathcal{N}^*(R) = R_j \tag{III.19}$$

ensuite nous avons

$$\text{Sim}(R_i, R_j) < \text{Sim}(R, R_j) \implies \mathcal{N}^*(R_i) = \emptyset \tag{III.20}$$

où $\mathcal{N}^*(\cdot)$ indique le meilleur voisin de région.

3. Troisièmement, nous assurons la priorité entre les agrégations de régions en validant uniquement les agrégations préférentielles. Une agrégation est préférentielle lorsque les régions ont une similarité supérieure à $\mathcal{S}_{it}$.

5.4. Seuil de similarité adaptatif

Dans les techniques classiques de segmentation d'images par croissance de régions, les régions comparées appartiennent à la même itération de fusion. Il n'y a pas de considérations simultanées de régions générées à partir de différentes itérations. Cependant, comme la taille et le contenu des régions changent à travers les itérations, une valeur de similarité fixe peut ne pas exprimer la même similarité visuelle à différentes itérations. Par exemple, si l'on considère la similarité de couleur comme critère de fusion, une valeur de similarité donnée ne correspondra pas à la même similarité visuelle lorsqu'on compare des petites régions homogènes (itérations précoces) et lorsque des grandes régions hétérogènes (après quelques itérations) sont comparées. Cela signifie que la valeur du critère de fusion ne doit pas être une valeur fixe, mais doit être mise à jour pour rester cohérente à travers les itérations. Les auteurs de Baatz et Shäpe (2000) ont introduit cette dimension dans le critère de fusion en comparant les régions candidates à la fusion avec la région résultante. Tandis que cette approche permet de partiellement corriger ce désavantage car elle se limite à deux itérations successives.

Cette idée est introduite dans l'approche proposée en utilisant un seuil de similarité adaptatif qui est mis à jour en fonction des résultats de l'itération précédente. Essentiellement, ce seuil est diminué lorsqu'il n'y a pas eu d'agrégations à l'itération précédente, sinon il est augmenté. Dans le cas où aucune agrégation ne s'est produite à l'itération précédente, le seuil de similarité est trop élevé, il est donc diminué pour permettre les agrégations à l'itération suivante. Dans le cas contraire, la similarité des régions nouvellement formées avec leur voisinage augmentera à mesure que leur contenu rassemblera le contenu des deux régions qui les ont formées. Ainsi, le seuil est augmenté afin de filtrer les régions voisines non pertinentes. Le seuil de similarité est mis à jour en fonction du coefficient d'itération α_{it}. En appliquant ce seuil de similarité adaptatif, nous assurons que les meilleures agrégations de régions arrivent en premier, ce qui nous permet d'imposer un ordre de priorité à ces agrégations durant le processus de croissance de régions. Nous utilisons

une formule de pénalisation de similarité proposée dans (Yu et Clausi, 2008).

$$\alpha_{it} = exp- \left(\frac{C_r}{\Delta * M_r} \right)^2 \tag{III.21}$$

où $C_r = \|candidateRegions_{i-1}\|$ est le nombre de régions candidates à la fusion, $M_r = \|mergedRegions_{i-1}\|$ donne le nombre de régions qui ont été fusionnées, $\|\cdot\|$ indique la taille de l'ensemble considéré et i est le rang de l'itération en cours. Δ est un paramètre qui indique la pente de la courbe.

6. Évaluation expérimentale

Pour valider l'approche proposée, nous l'avons appliqué sur la base d'images BSDS500 de Berkeley (Martin *et al.*, 2001) et celle de Weizmann (Sharon *et al.*, 2007). BSDS500 est une base qui est constituée de `500` images naturelles avec `5` différentes segmentations réalisées par des humains (vérité terrain - VT) pour chaque image. L'approche proposée est comparée à d'autres approches de segmentation d'image sur `100` images, sélectionnées au hasard à partir de BSDS500.

La base de Weizmann contient `200` images de niveau de gris accompagnées de leurs segmentations de vérité terrain. La base est spécialement conçue pour éviter les ambiguïtés potentielles en incorporant uniquement des images qui représentent clairement un ou deux objets au premier plan qui diffèrent de leur environnement par leur intensité, leur texture ou tout autre indice de bas niveau. La VT a été obtenue par segmentation manuelle par des sujets humains en deux ou trois classes. Chaque image est segmentée par trois personnes différentes.

Comme la similarité globale entre deux régions est normalisée, nous établissons empiriquement la valeur de la similarité d'arrêt $\mathcal{S}_0$ à `0,4` durant toutes les expériences. Les principaux résultats sont présentés et discutés dans la suite de chapitre.

6.1. Critères d'évaluation

Soit $S = \{S_j\}$ la partition résultat obtenue après application d'une technique de segmentation sur une image $\mathcal{I}$ et $G = \{G_i\}$ la VT correspondante, nous exprimons par $TP(G,S)$ et $FN(G,S)$ respectivement le nombre total de pixels de contours vrais-positifs et faux-négatifs de S par rapport à G. Aussi, l'image $\mathcal{I}$ est composée de N pixels. Nous considérons alors les six critères d'évaluation de la qualité d'une segmentation suivants pour comparer l'approche proposée à quelques méthodes de l'état-de-l'art. Le Boundary Recall (BR) et le Undersegmentation Error (UE) sont utilisés pour évaluer les résultats de la sur-segmentation CoSLIC, tandis que les quatre derniers critères sont utilisés pour évaluer les résultats de la segmentation finale.

1. le **BR** mesure la fraction des contours de VT qui se situent dans au moins une limite de superpixel, avec une distance de tolérance δ (généralement fixée à 2 pixels). Il est défini par :

$$BR(G,S) = \frac{TP(G,S)}{TP(G,S) + FN(G,S)} \tag{III.22}$$

2. l'**UE** compare les zones de superpixels pour mesurer dans quelle mesure les superpixels inondent les contours de la région VT. Il existe dans la littérature au moins trois formules (Achanta *et al.*, 2012 ; Levinshtein *et al.*, 2009 ; Neubert et Protzel, 2012) pour cette mesure. Nous utilisons dans la suite du manuscrit celle proposée par (Neubert et Protzel, 2012) car, par rapport aux autres, elle permet une

meilleure évaluation de l'adhérence aux contours par les segmentations obtenues. Sa formule est comme suit :

$$\mathbf{UE}(G,S) = \frac{1}{N}\sum_{G_i}\sum_{S_j\cap G_i\neq\emptyset} min\Big\{|S_j \cap G_i|,\ |Sj - G_i|\Big\} \tag{III.23}$$

3. Probabilistic Rand Index (**PRI**) (Unnikrishnan *et al.*, 2007) est une mesure de la similarité entre deux groupes de données en fonction de leur étiquetage. En segmentation d'image, il mesure la proportion de pixels ayant les mêmes étiquettes que la segmentation VT.

$$\mathbf{PRI}(S,G) = \frac{1}{N}\sum_{i<j}\Big[c_{ij}\times p_{ij} + (1-c_{ij})\times(1-p_{ij})\Big] \tag{III.24}$$

où c_{ij} indique que les pixels i et j ont la même étiquette et p_{ij} sa probabilité.

4. Variation of Information (**VoI**) (Meilă, 2007) donne la dissemblance entre deux résultats de regroupement. Sur la base d'une expression d'entropie conditionnelle, VoI mesure la quantité de pixels regroupés de manière aléatoire dans la segmentation, sans aucun indice dans le VT.

$$\mathbf{VoI}(G,S) = H(G) + H(S) + 2\times I(G,S) \tag{III.25}$$

avec H et I respectivement l'entropie et l'information mutuelle entre les deux segmentations G et S.

5. Boundary Displacement Error (**BDE**) (Freixenet *et al.*, 2002) mesure l'erreur de déplacement moyenne d'un pixel de contour et les pixels contours les plus proches dans la segmentation VT.

$$\begin{aligned}\mathbf{BDE}(G,S) &= \frac{1}{2}\Big[D_G^S + D_S^G\Big];\\ D_{B_1}^{B_2} &= \frac{1}{N}\sum_{x\in B_1} min\{d_E(x,y)\}_{\forall y\in B_2}\end{aligned} \tag{III.26}$$

où $D_{B_1}^{B_2}$ mesure la distance entre les deux images de contours B_1 et B_2. La fonction $d_E(\cdot,\cdot)$ est la distance Euclidienne entre deux points.

6. Global Consistency Error (Martin *et al.*, 2001) (**GCE**) mesure la possibilité dans laquelle une segmentation est un raffinement de l'autre.

$$\begin{aligned}\mathbf{GCE}(G,S) =& \frac{1}{N}\ min\Big\{\sum_i E(G,S,p_i),\ \sum_i E(S,G,p_i)\Big\};\\ E(S_1,S_2,p_i) =& \frac{|R(S_1,p_i)\setminus R(S_2,p_i)|}{|R(S_1,p_i)|}\end{aligned} \tag{III.27}$$

où $R(S,p)$ est l'ensemble de pixels correspond à la région de la segmentation S contenant le pixel p.

Évidemment, plus la valeur de BR est élevée, meilleure est la sur-segmentation, contrairement à l'UE, dont les valeurs élevées indiquent une qualité de sur-segmentation faible. Un bon algorithme de segmentation devrait avoir des valeurs PRI élevées, ainsi que de faibles VoI, BDE et GCE.

Sachant que plusieurs VT sont fournies pour chaque image, nos résultats de segmentation sont évalués en comparant chaque image segmentée avec toutes les images VT séparément. La valeur du critère d'évaluation

résultante pour une image est la moyenne de la comparaison par rapport aux différentes VT.

6.2. Décomposition en superpixels par CoSLIC

Le but de ce travail est de segmenter les images à travers la croissance de la région basée sur le superpixel. Le tableau III.2 récapitule les performances des deux algorithmes de décomposition en superpixels sur les base d'images BSDS500 et Weizmann. L'allure de ces résultats est graphiquement présentée dans la figure III.6. Notons qu'en global l'algorithme CoSLIC produit de meilleurs résultats que le SLIC, comme le montre quelques exemples présentés dans la figure III.7. Plus particulièrement, le CoSLIC surclasse le SLIC dans les images contenant des objets voisins séparés par des contours fins. Notamment, en termes de BR, les performances de notre algorithme de décomposition en superpixels proposé sont supérieures à celles du SLIC sur les deux bases d'évaluation. Ce résultat est assez prévisible car CoSLIC est conçu pour offrir une meilleure adhérence aux contours par rapport au SLIC. Cependant, le SLIC présente de meilleurs résultats en termes d'UE sur les images de Weizmann tandis que le CoSLIC l'emporte sur celle de BSDS500. La figure III.7 présente quelques résultats du CoSLIC et du SLIC ainsi que les images de la segmentation finale résultantes. L'intérêt de la correction par contours globaux ajoutée apparaît clairement. Les résultats de segmentation d'une décomposition initiale avec SLIC (III.7-d) et de CoSLIC (III.7-e) montrent des changements radicaux sur les résultats finaux lorsque quelques contours ne sont pas détectés par l'algorithme SLIC en phase de décomposition en superpixels.

TABLE III.2. – Résultats d'évaluation de la décomposition d'image en superpixels par le SLIC et le CoSLIC sur les bases d'images BSDS500 et Weizmann. Les meilleurs résultats sont en **gras**.

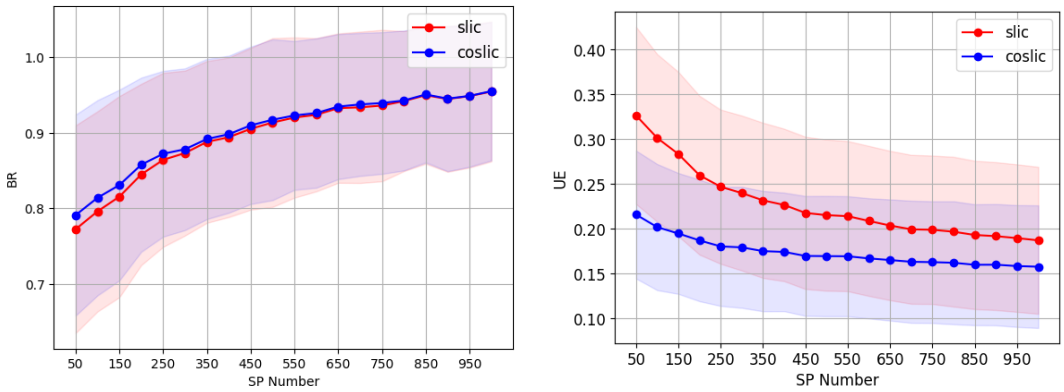
		Boundary Recall ↑			**Under-segmentation Error ↓**		
		Min	**Max**	**Moy**	**Min**	**Max**	**Moy**
BSDS500	SLIC	0.772	**0.955**	0.898	0.187	0.327	0.227
	CoSLIC	**0.791**	**0.955**	**0.903**	**0.158**	**0.216**	**0.174**
Weizmann	SLIC	0.781	0.949	0.895	**0.153**	**0.372**	**0.278**
	CoSLIC	**0.809**	**0.952**	**0.906**	0.237	0.386	0.320

6.3. Segmentation par agrégations adaptatives

À partir de la décomposition en superpixels de l'image, la segmentation est obtenue par agrégation successives de superpixels voisins selon le principe exposé dans la section 5. Les résultats de la segmentation sont évalués par comparaison avec les segmentations VT associées en termes de cohérence du contenu et le degré de fragmentation.

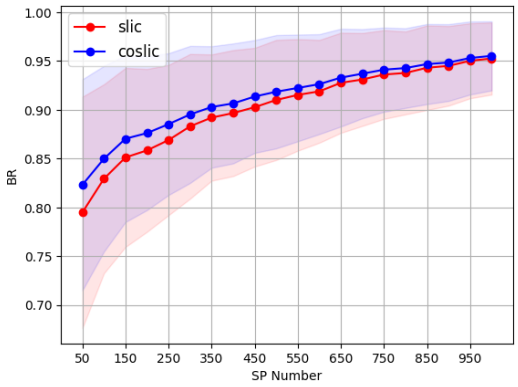
6.3.1. Similarité de superpixels

Ce chapitre présente une mesure de similarité entre régions, énoncée dans la section 5.1, qui est formée d'une composante frontière AdRG-Sim_B et d'une composante contenu AdRG-Sim_C. Le tableau III.3 présente les résultats d'évaluation de ces différentes composantes de la mesure de similarité. La figure III.7

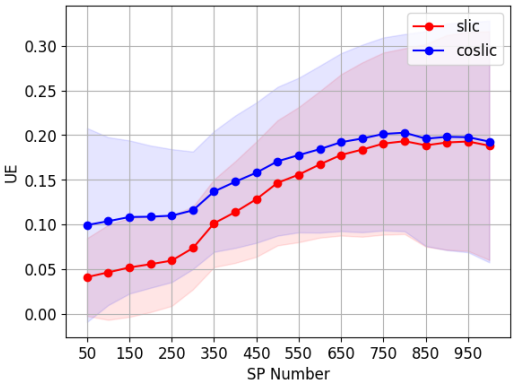


BR BSDS500

UE BSDS500



BR Weizmann



UE Weizmann

FIGURE III.6. – Évaluation des résultats de la décomposition en superpixels : courbes de BR (III.6a & III.6c) et de UE (III.6b & III.6d) pour SLIC et CoSLIC sur les jeux de données BSDS500 et Weizmann.

montre une comparaison graphique de ces résultats.

Dans leur ensemble, ces résultats montre que la composante de frontière produit les meilleurs résultats vis-à-vis du critère GCE. Ce comportement corrobore le principe de cette composante qui est axé sur la différence au niveau des frontières. Cette tendance est aussi confirmée par le graphe III.7h de la figure III.7 où les valeurs moyenne et médiane sont beaucoup plus faibles que celles des autres composantes pour les images de la base Weizmann. Cette différence est moins prononcée sur la base d'images BSDS500(III.8d). La composante de contenu présente ses meilleures performances en termes de GCE sur les deux base d'images et des résultats moyens en termes de PRI. Ces derniers résultats peuvent montrer les cas difficiles pour cette mesure. La dominance en GCE montre bien qu'avec cette composante les régions formées possèdent la meilleure cohérence. Les meilleurs résultats en termes de VoI sont obtenus sur la base Weizmann par la mesure composite AdRG-*Sim*. Elle offre, dans les reste des critères, des résultats entre ses deux composantes. Toutefois en considérant tous les critères, cette dernière mesure fournit le meilleur équilibre entre la détection des contours et la cohérence des régions formées.

6.3.2. Stratégie de fusion de superpixels

À chaque itération, La stratégie de fusion proposée permet de choisir les paires de superpixels à fusionner. Notamment, elle assure une sélection locale optimale pour les superpixels tout en évitant les conflits au niveau global de l'image. Par ailleurs, les fusions sont cadencées par une procédure de filtrage adaptatif

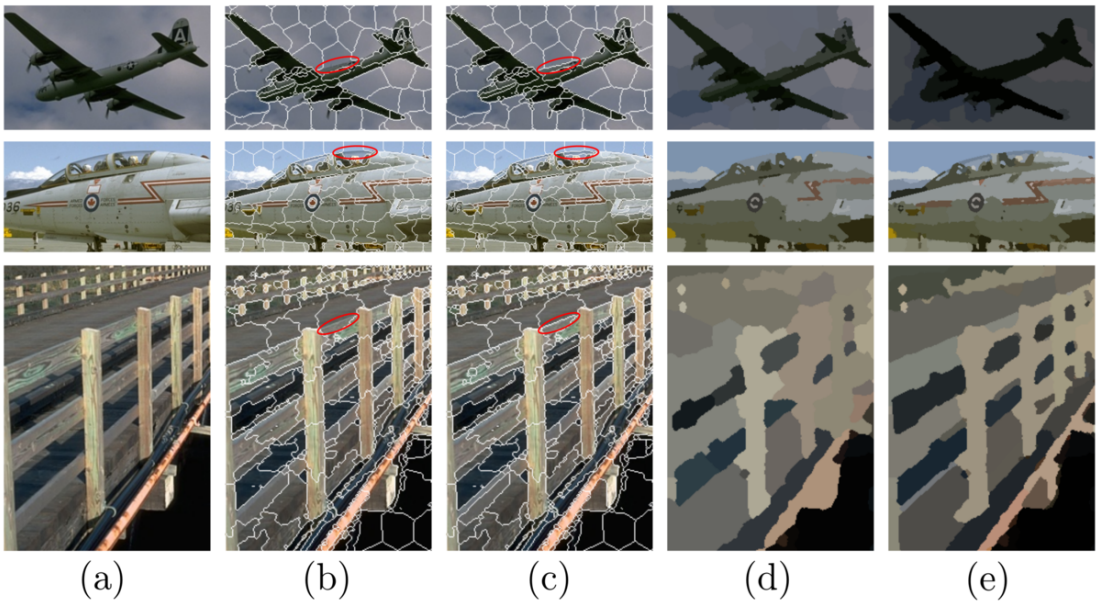
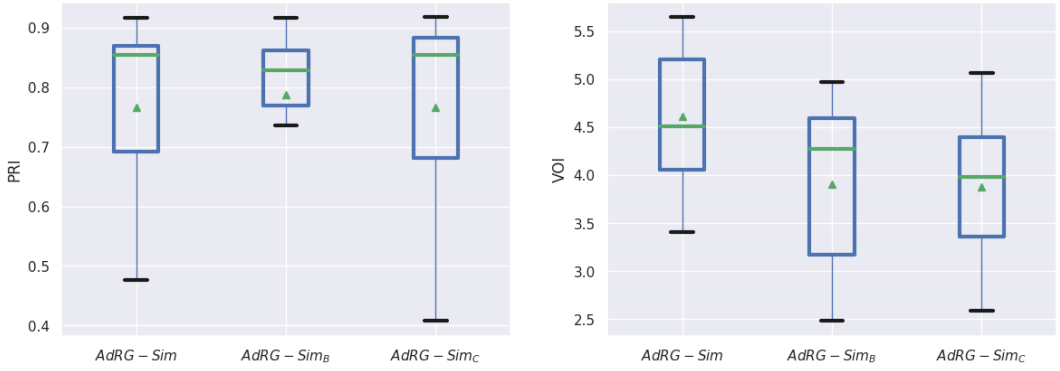


FIGURE III.7. – Quelques résultats de l'extension proposée du SLIC en utilisant des contours globaux pour améliorer l'adhérence aux contours. (a) image originale. Les superpixels SLIC sont donnés en (b) et le résultat corrigé en utilisant les contours globaux est fourni par (c). Les contours globaux permettent de récupérer certains contours omis par le SLIC. (d) et (e) donnent des résultats de segmentation par croissance de régions, respectivement à partir des superpixels SLIC et CoSLIC.

TABLE III.3. – Résumé des résultats de l'évaluation des performances de notre approche (AdRG), NCut, CTM, HFEM, VoI-BFM et MeanShift sur BSDS500 et Weizmann. La meilleure segmentation devrait avoir le PRI le plus élevé, tandis que la VoI, le BDE et la GCE devraient être les plus bas. Les meilleurs résultats sont écrits en **gras**.

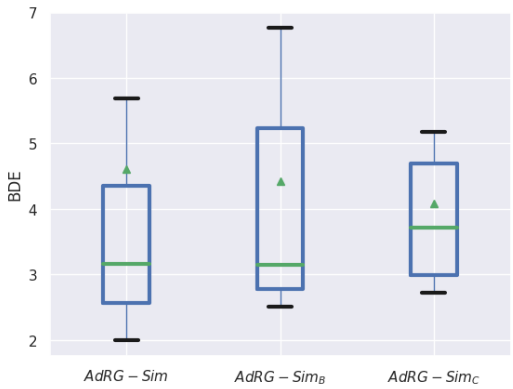
		PRI ↑			VoI ↓			BDE ↓			GCE ↓		
		Min	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**
BSDS500	AdRG-Sim_C	0.408	**0.918**	0.766	2.586	5.063	**3.881**	2.723	**7.945**	**4.085**	0.044	**0.475**	**0.324**
	AdRG-Sim_B	0.380	0.917	**0.787**	**2.479**	**4.973**	3.908	2.513	9.064	4.429	**0.036**	0.502	0.325
	AdRG-Sim	**0.476**	0.917	0.767	3.406	5.647	4.608	**1.996**	17.669	4.612	0.096	0.599	0.401
Weizmann	AdRG-Sim_C	0.402	0.930	0.700	0.811	5.164	2.193	**4.108**	19.146	11.025	**0.025**	**0.373**	**0.122**
	AdRG-Sim_B	**0.461**	0.926	**0.729**	0.591	5.158	**2.148**	4.176	**18.714**	**10.047**	0.048	0.445	0.146
	AdRG-Sim	0.257	**0.961**	0.680	**0.469**	**5.034**	2.322	4.154	18.798	11.251	0.035	0.473	0.151

pour favoriser la cohérence des régions résultantes. Le tableau III.4 et la figure III.7 résument les résultats de l'évaluation de l'effet de l'aspect adaptatif de l'approche proposée. Ces résultats montrent, sur l'ensemble des critères d'évaluation, que la segmentation adaptative AdRG-Sim_{wAd} apporte un gain notoire sur les performances de l'algorithme. Plus particulièrement, les rectangles plus restreints en PRI et VoI pour la segmentation non-adaptative AdRG-*Sim* affichés dans la figure III.7 indiquent des valeurs contiguës mais AdRG-Sim_{wAd} exhibe les meilleures valeurs moyennes et médianes. En termes de BDE et GCE les

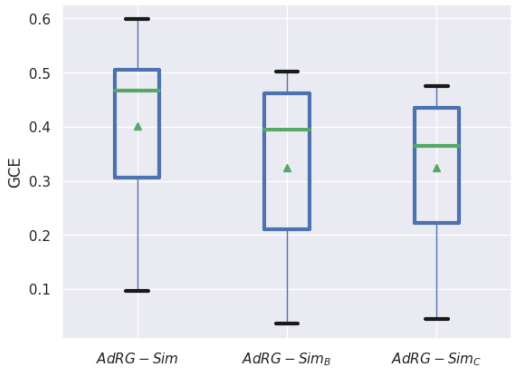


PRI - BSDS500

VoI - BSDS500



BDE - BSDS500



GCE - BSDS500

rectangles ainsi que les valeurs centrales de l'approche adaptative AdRG-Sim_{wAd} sont les meilleures.

TABLE III.4. – Résumé des résultats de l'évaluation des performances de notre approche (AdRG), NCut, CTM, HFEM, VoI-BFM et MeanShift sur BSDS500 et Weizmann. La meilleure segmentation devrait avoir le PRI le plus élevé, tandis que la VoI, le BDE et la GCE devraient être les plus bas. Les meilleurs résultats sont écrits en **gras**.

		PRI ↑			VoI ↓			BDE ↓			GCE ↓		
		Min	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**
BSDS500	AdRG-*Sim*	0.476	0.917	0.767	6.204	7.748	7.125	3.107	**5.085**	**4.058**	0.430	0.710	0.621
	AdRG-Sim_{wAd}	**0.767**	**0.930**	**0.874**	**3.406**	**5.647**	**4.608**	**1.996**	17.669	4.612	**0.096**	**0.599**	**0.401**
Weizmann	AdRG-*Sim*	0.205	0.646	0.409	4.867	7.800	6.405	6.737	49.143	20.504	0.064	**0.465**	0.193
	AdRG-Sim_{wAd}	**0.257**	**0.961**	**0.680**	**0.469**	**5.034**	**2.322**	**4.154**	**18.798**	**11.251**	**0.035**	0.473	**0.151**

6.3.3. Comparaison aux travaux similaires

Pour évaluer l'approche proposée, les résultats obtenus sont comparés à ceux de quatre approches de segmentation d'images bien connues de la littérature : Normalized Cut (NCut) (Shi et Malik, 2000), MeanShift (Comaniciu et Meer, 2002), CTM (Yang *et al.*, 2008), et HFEM (Yin *et al.*, 2017). Dans NCut,

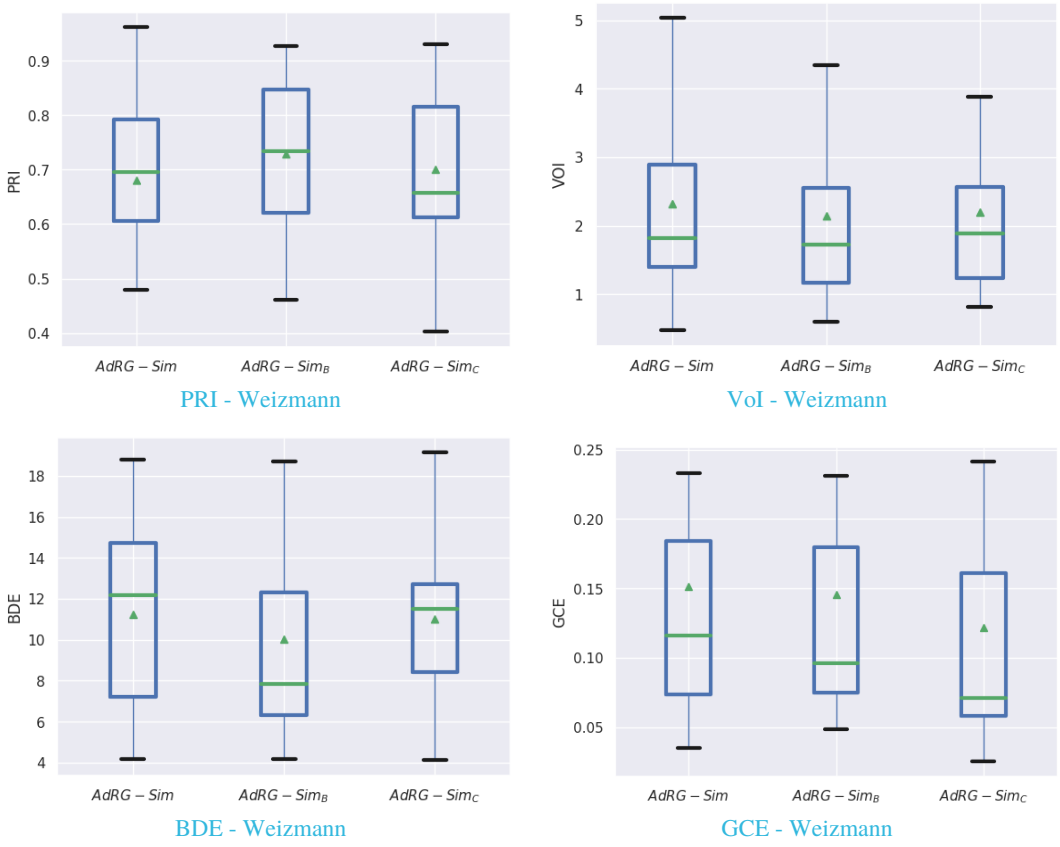
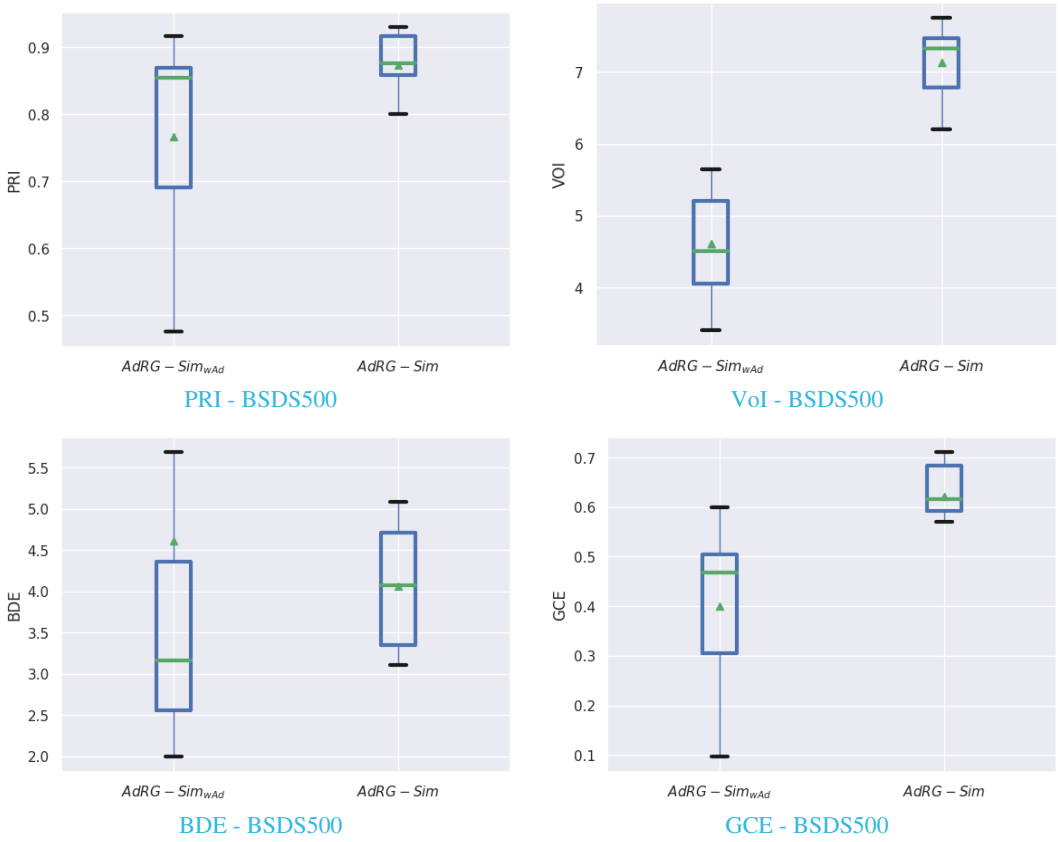


FIGURE III.7. – Évaluation des résultats des résultats de la segmentation en fonction de la composante de mesure de similarité utilisée sur les bases d'images BSDS500 et Weizmann.

une image est modélisée par un graphe. La segmentation est obtenue en partitionnant le graphe tout en minimisant les coupures globales. Cette approche met l'accent sur la similarité entre les segments d'image. L'approche du MeanShift est basée sur l'application d'un estimateur de densité non paramétrique à la segmentation de l'image. Les pixels adjacents sont regroupés selon un critère de similarité pour former les régions de segmentation finale. L'approche CTM transforme la segmentation d'images naturelles en un problème de regroupement de données de texture mixte multivarié. Les caractéristiques de texture sont calculées à l'aide de bancs de filtres de texture bidimensionnelles ou de fenêtres simples de taille fixe. Les auteurs montrent que les mixtures de textures peuvent être efficacement segmentées par un algorithme de regroupement agglomératif simple dérivé d'une approche de compression de données avec perte en minimisant la longueur de codage globale des vecteurs de caractéristiques. Le HFEM (ou segmentation hiérarchique non supervisée par maximisation de l'entropie floue) est un algorithme de segmentation multi-niveau non supervisé. L'algorithme maximise à la fois une bipartition floue de l'image et la régularité de la segmentation grâce à un opérateur de segmentation à deux niveaux qui utilise des coupes graphiques binaires.

Dans la tâche de segmentation, nous nous concentrons particulièrement sur la mesure de similarité et la stratégie de fusion dans le processus de croissance de régions. La stratégie de fusion proposée est conçue pour guider les itérations de fusion pendant la segmentation. Cette stratégie parvient effectivement à fusionner d'abord des voisins visuellement similaires, comme illustré par la figure III.8 au fur des itérations.



PRI - BSDS500 VoI - BSDS500

BDE - BSDS500 GCE - BSDS500

Au vu des résultats présentés dans le tableau III.5 et la figure III.7, l'approche proposée produit dans l'ensemble des résultats compétitifs par rapport aux approches comparées. Par ailleurs, nous notons que l'approche proposée fournit le meilleur résultat en termes de PRI et BDE. En plus des résultats visuels fournis dans la figure III.9, cela confirme la qualité des superpixels initiaux mais aussi de la mesure de la similarité proposée. Cependant, dans certains cas, comme on peut le voir sur les images présentées dans la figure III.9-image 6, notre approche tend à sur-segmenter les images. Ce comportement peut être dû à la valeur de similarité finale qui arrête les agrégations de régions. Cette valeur est déterminée empiriquement et peut donc ne pas convenir à toutes les images.

Les performances selon les critères PRI ont tendance à conclure que notre approche est supérieure à NCut et vient légèrement en dessous des autres approches. Ce critère mesure la qualité de la classification des pixels par rapport à la VT. En effet, l'approche proposée n'inclut pas explicitement de contraintes sur l'étiquetage des pixels mais gère tout cela via l'ordre d'agrégation introduit dans la stratégie de fusion(§5.3). Le HFEM a les meilleurs résultats à la fois en VoI et en GCE. L'approche proposée surpasse le MeanShift en termes de GCE et donne de meilleurs résultats que MeanShift et CTM en termes de VoI. En fait, notre approche, contrairement aux autres approches, ne considère pas une vue globale au niveau de l'image mais uniquement au niveau de la région.

7. Conclusions

Ce chapitre a décris une approche de segmentation d'image basée sur une agrégation itérative de régions qui présente trois contributions principales. Tout d'abord, une extension pour l'algorithme SLIC est proposée afin de fournir une meilleure adhérence aux contours. Deuxièmement, nous avons proposé

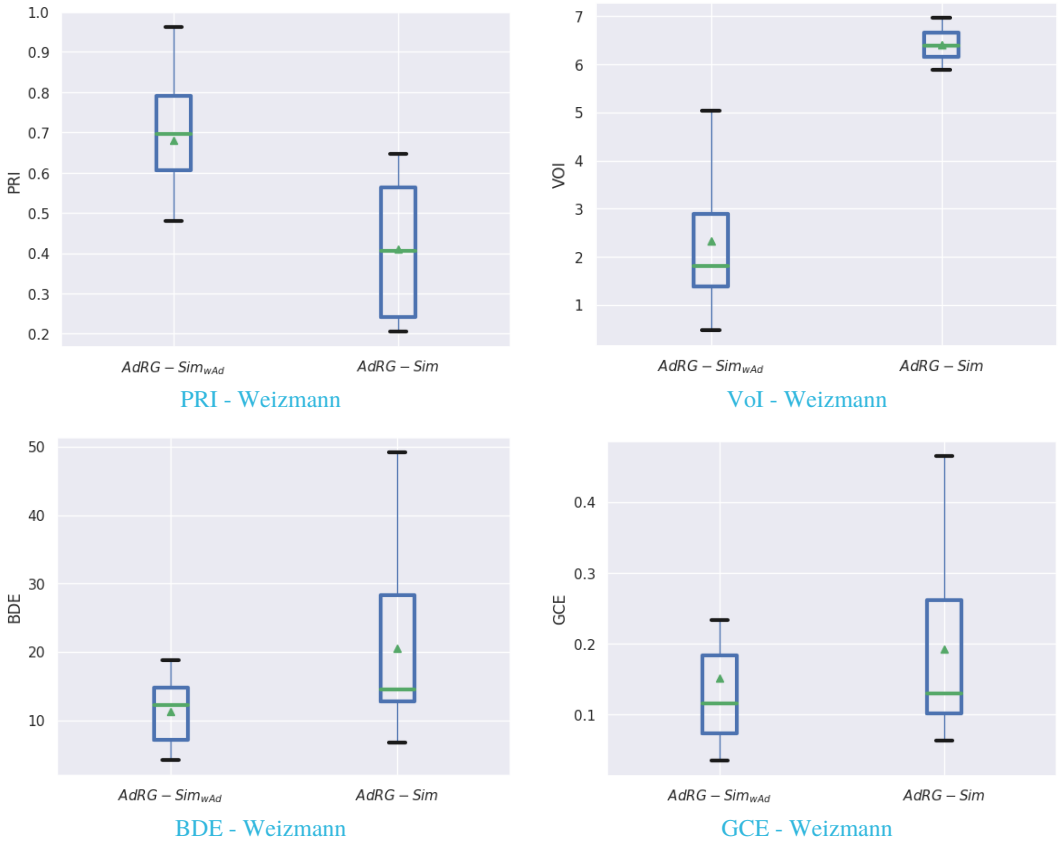


FIGURE III.7. – Évaluation des résultats de la segmentation par agrégations adaptatives sur les images de BSDS500 et Weizmann.

TABLE III.5. – Résumé des résultats de l'évaluation des performances de notre approche (AdRG), NCut, CTM, HFEM, VoI-BFM et MeanShift sur BSDS500 et Weizmann. La meilleure segmentation devrait avoir le PRI le plus élevé, tandis que la VoI, le BDE et la GCE devraient être les plus bas. Les meilleurs résultats sont écrits en **gras**.

		PRI ↑	VoI ↓	BDE ↓	GCE ↓
BSDS500	NCut (Shi et Malik, 2000)	0.739	2.913	17.156	0.223
	MeanShift(Comaniciu et Meer, 2002)	0.777	4.317	13.161	0.581
	CTM (Yang *et al.*, 2008)	**0.779**	6.219	19.198	0.365
	HFEM (Yin *et al.*, 2017)	0.777	**2.307**	10.670	**0.2215**
	AdRG-Sim_{wAd}	0.763	3.804	**10.160**	0.448
Weizmann	NCut (Shi et Malik, 2000)	0.652	2.217	12.298	0.167
	MeanShift(Comaniciu et Meer, 2002)	0.485	3.001	13.508	0.155
	VoI-BFM (Mignotte, 2014)	0.607	**2.112**	14.156	**0.091**
	AdRG-Sim_{wAd}	**0.680**	2.322	**11.251**	0.151

une mesure de similarité robuste pour comparer des régions en utilisant à la fois des critères globaux

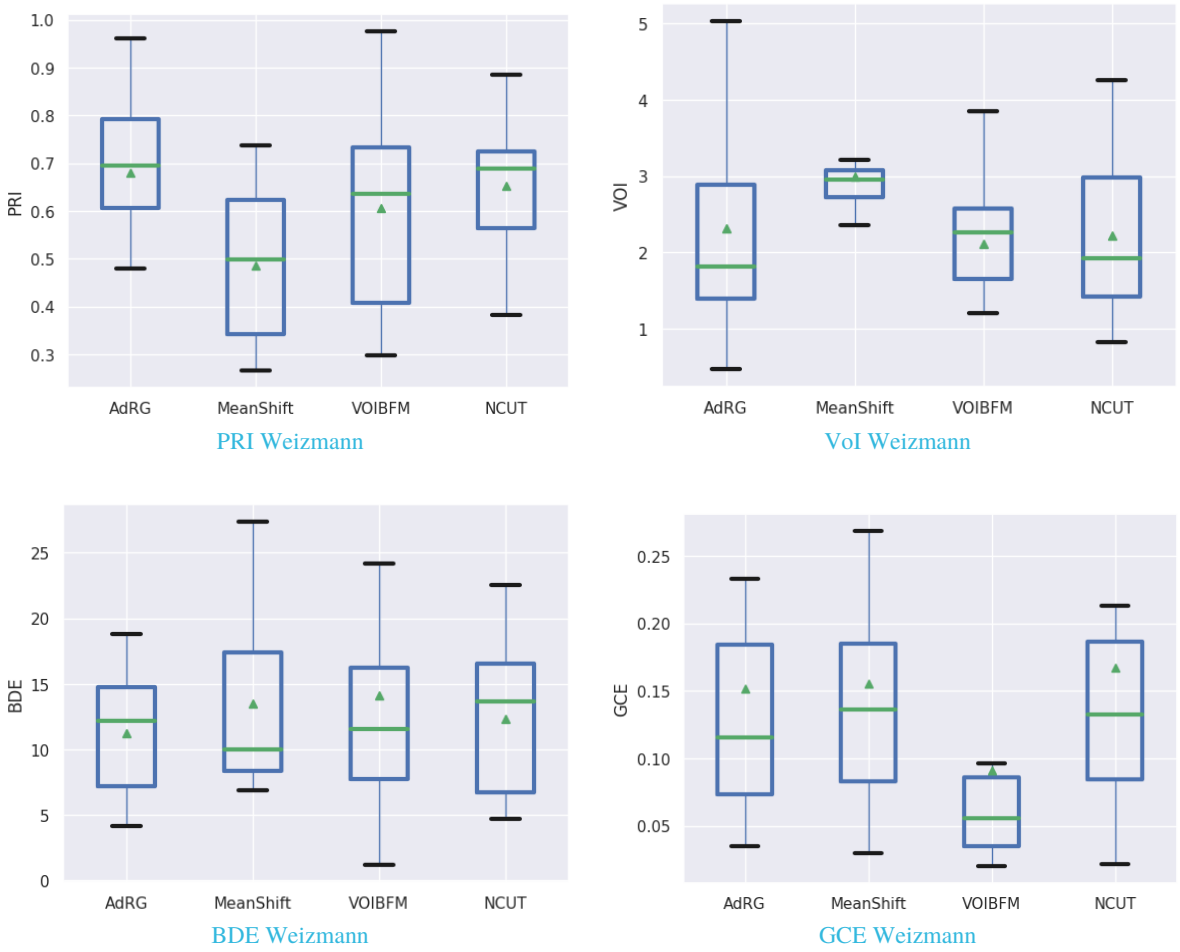


FIGURE III.7. – Évaluation des résultats de la segmentation : courbes de PRI, VoI, BDE et GCE sur les bases d'images Weizmann.

multi-échelles et une comparaison locale à la frontière commune. Cette mesure intègre des informations sur les frontières pour éviter la fusion d'objets superposés. Troisièmement, une stratégie de fusion est conçue pour contrôler les agrégations de régions au fur des itérations en utilisant un seuil de similarité adaptatif. Cette stratégie garantit que les agrégations se produisent dans l'ordre de similarité décroissante. L'efficacité des contributions proposées a été évaluée par des comparaisons avec des approches de segmentation bien établies sur la base d'image BSDS500.

Bien que produisant des résultats encourageant, la mesure de similarité proposée ne permet pas de regrouper des régions visuellement différentes même lorsque celle-ci appartiennent au même objet. Ce problème peut d'ailleurs être généralisé à toutes les mesures classiques qui sont définies sur les descripteurs des régions. Le chapitre suivant propose une autre approche de comparaison permettant de combler cette lacune.

FIGURE III.8. – Processus de segmentation : quelques résultats intermédiaires de la segmentation.

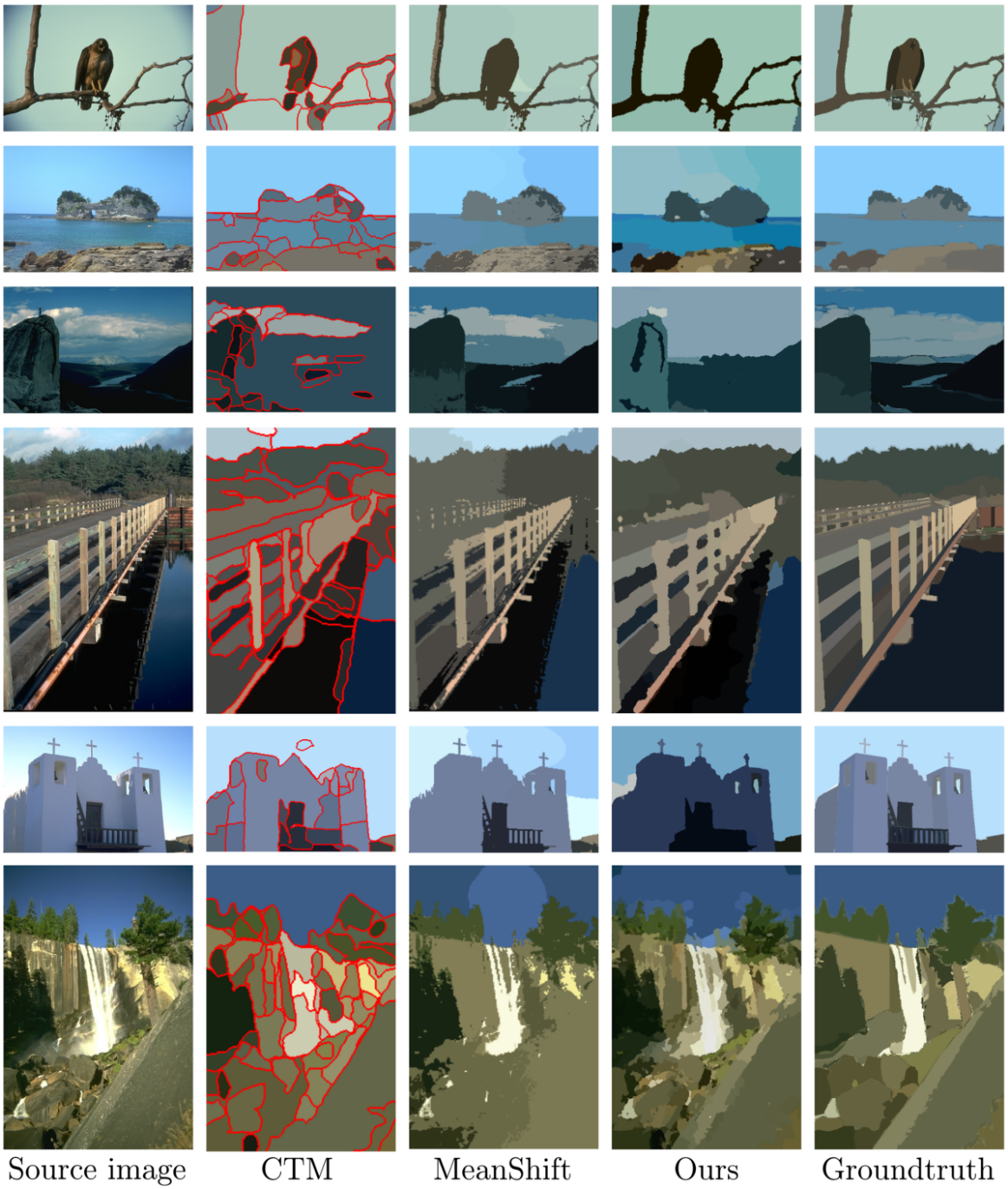


FIGURE III.9. – Résultats de la segmentation : quelques résultats de notre approche de segmentation avec ceux de CTM et MeanShift sur la base d'image BSDS500.

IV. Similarité des superpixels par apprentissage

À l'instar du chapitre précédent, ce chapitre traite de la thématique de comparaison de régions dans une approche de segmentation par agrégations itératives de régions à partir d'une décomposition initiale en superpixels. À l'opposé de ce dernier chapitre où les caractéristiques des superpixels sont extraites par des fonction calculées, ici ces dernières sont estimées grâce à un apprentissage automatique. La contribution principale de cette partie est l'utilisation de forêts aléatoires pour apprendre la probabilité de fusion entre superpixels adjacents. Par rapport aux travaux existants, cette approche basée sur l'apprentissage apprend les règles de fusion sans calcul de mesure de similarité explicite. Cela fournit un processus simplifié de fusion de régions qui est appliqué de manière itérative pour produire une segmentation des images. L'utilisation de l'apprentissage permet d'éviter le fossé sémantique entre les superpixels à comparer.

Sommaire

1. Introduction

En interprétation automatique d'images, le fossé sémantique est défini comme le gap entre la représentation bas niveau et l'interprétation haut niveau du contenu (Smeulders *et al.*, 2000) faite par l'humain. Plusieurs facteurs contribuent à cet écart sémantique parmi lesquels, particulièrement dans un contexte de segmentation par croissance de région le calcul de similarité entre régions est l'un de ces facteurs à cause d'au moins deux raisons.

— la mesure de similarité est intrinsèquement dépendant des descripteurs utilisés pour caractériser ces régions. Ainsi la similarité entre deux régions sera influencée par les défauts des descripteurs des régions, sachant qu'il n'existe pas de descripteur parfait.

— la sémantique de la mesure de similarité est définie par l'allure de la formule mathématique utilisée pour calculer la degré de ressemblance inter-régions.

Ce chapitre aborde la problématique du fossé sémantique dans le calcul de la similarité entre régions pour une approche de segmentation. Nous présentons une approche de segmentation qui utilise des superpixels comme entrées. Plutôt que de calculer une mesure de similarité pour comparer des superpixels, nous proposons d'entrainer un classifieur ML pour déduire la probabilité de fusion de deux superpixels. Cela permettra notamment de fusionner en toute transparence des superpixels visuellement différents appartenant au même objet sémantique. Dans ce travail, nous avons choisi la méthode des forêts aléatoires pour apprendre la fusion des superpixels. Ce choix est basé sur plusieurs propriétés de la technique ML dont les principales sont l'efficacité de calcul, la robustesse par rapport aux valeurs aberrantes et les résultats probabilistes.

2. Croissance de régions par apprentissage

L'approche de segmentation proposée dans ce chapitre utilise un classifieur RF pour estimer la similarité entre les superpixels. En outre, en représentant les superpixels de l'image à l'aide d'un graphe d'adjacence de régions (GAR), la segmentation est réalisée par agrégations itératives des nuds du graphe.

Au préalable, nous supposons que l'ensemble de toutes les images de la base, noté $\mathcal{D}$, est composé d'images représentant une scène unique qui est formée de K classes d'objets, $\{C_1, C_2, \ldots, C_K\}$. En outre, $\mathcal{D}$ est divisé en deux sous-ensembles : un ensemble d'apprentissage $\mathcal{D}_{train}$ et un ensemble de tests $\mathcal{D}_{test}$. Le sous-ensemble $\mathcal{D}_{train}$ est formé d'une sélection aléatoire d'images de $\mathcal{D}$ chacune accompagnées de son image d'annotations vérité terrain (VT) alors que $\mathcal{D}_{test} = \mathcal{D} \setminus \mathcal{D}_{train}$ contient les images à segmenter. Un exemple d'un tel ensemble de données est une séquence vidéo où un sous-ensemble aléatoire des trames est fourni avec des annotations VT. L'approche de segmentation proposée ici se base sur la technique d'apprentissage automatique pour construire d'abord un modèle de fusion superpixels à partir de $\mathcal{D}_{train}$, puis effectue une segmentation des images à partir de $\mathcal{D}_{test}$ par des agrégations récursives des superpixels. Cette approche est subséquemment composée de deux phases principales. Tout d'abord, une phase d'apprentissage d'un classifieur par RF (§ 3) est effectuée sur une décomposition initiale des superpixels des images de $\mathcal{D}_{train}$ afin d'apprendre comment fusionner les superpixels. En d'autres termes, le classifieur RF est entraîné à prédire si deux superpixels voisins quelconques appartiennent ou non au même objet sémantique. Deuxièmement, pour toute image donnée à partir de $\mathcal{D}_{test}$, une phase de segmentation (§ 6.2) est effectuée via des agrégations itératives de superpixels basées sur des prédictions faites par le classifieur RF entrainé dans la phase précédente. En particulier, l'utilisation du classifieur RF

permet la fusion des superpixels voisins similaires de manière itérative, au moyen de leur probabilité de fusion prédite. Cette étape itérative aboutit à une segmentation finale du contenu de l'image en des partitions homogènes. La figure IV.1 présente un aperçu de cette approche de segmentation proposée. En utilisant un

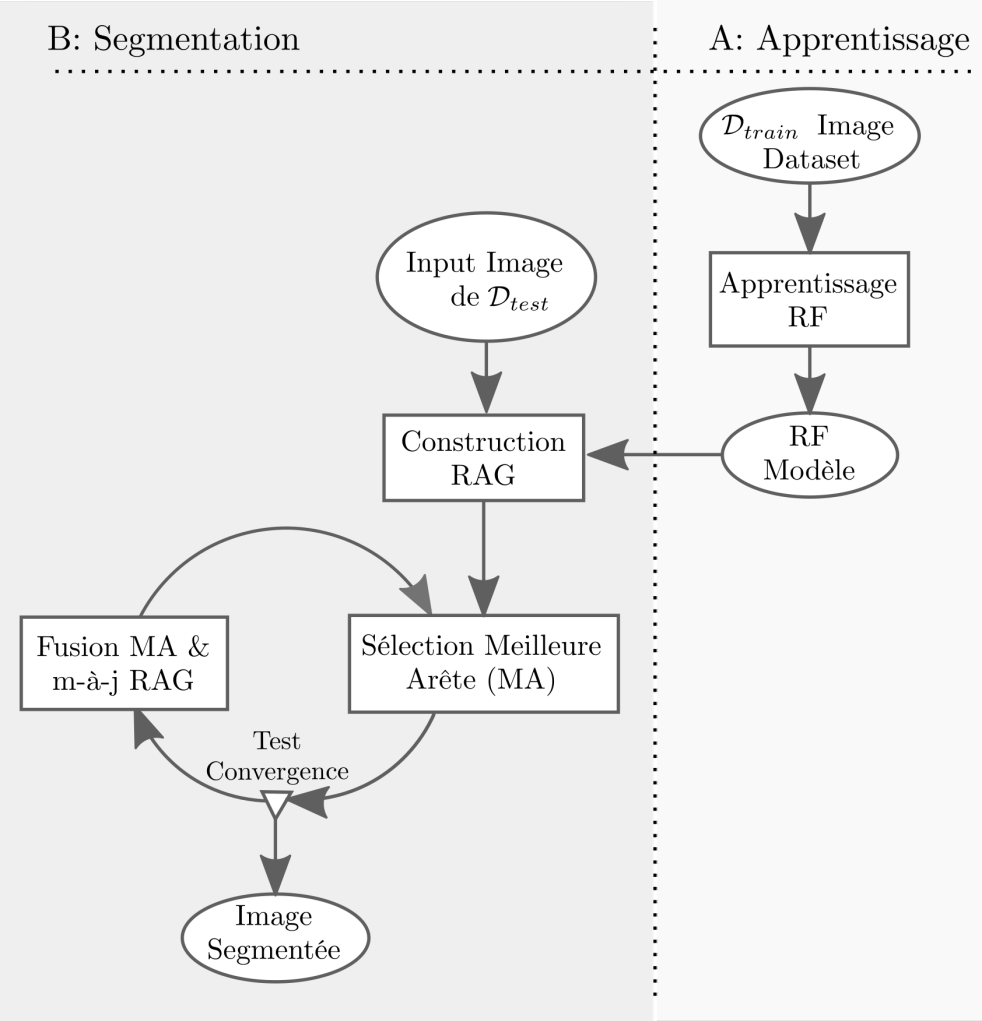


FIGURE IV.1. – Illustration de l'approche de segmentation. L'approche proposée commence par décomposition en superpixels de l'image source, puis construit un graphe d'adjacence de région (GAR) de ses superpixels. Ensuite, en utilisant les forêts aléatoires, la probabilité de fusion pour toutes les paires de superpixels voisins est calculée et des fusions itératives des meilleurs superpixels voisins sont réalisées pour finir avec une segmentation de l'image.

classifieur RF pour apprendre les agrégations de superpixels, la segmentation d'image proposée évite le fossé sémantique lié au calcul de la mesure de similarité dans les approches classiques de type croissance de régions. De plus, l'utilisation de superpixels au lieu de pixels permet une meilleure caractérisation des entités élémentaires manipulées et une réduction drastique du temps des calculs.

3. Apprentissage par forêts aléatoires (RF)

Les forêts aléatoires (Random Forests - RF) sont une technique d'apprentissage d'ensemble généralement utilisée pour les tâches de classification et de régression. Dans la communauté de l'apprentissage automatique (Machine Learning - ML), les RF ont été introduits par (Amit et Geman, 1997 ; Breiman, 2001). Plus

particulièrement, dans la communauté de la vision par ordinateur, leur popularité provient principalement des travaux de (Lepetit et Fua, 2006 ; Ozuysal *et al.*, 2007). Elles sont maintenant largement utilisées pour diverses tâches de classification supervisées, de la classification des images à la segmentation vidéo (Bosch *et al.*, 2007 ; Conze *et al.*, 2017 ; Yin *et al.*, 2007).

Il s'agit d'une collection d'arbres de décision non corrélés `T` dont chacun est construit à partir d'un sous-ensemble de `M` échantillons pris aléatoirement à partir des données d'apprentissage $\mathcal{X} = \{x_i,\ i = 1..\mathtt{N}\}$ où x_i décrit les données de l'échantillon i^{th}. A partir d'un sous-ensemble $\mathcal{X}_t$, l'arbre correspondant $\mathcal{T}_t$ est construit par bipartition récursive de $\mathcal{X}_t$ jusqu'à ce qu'à atteindre les conditions d'arrêt. En commençant par associer $\mathcal{X}_t$ à la racine de $\mathcal{T}_t$, l'algorithme génère, pour chaque nud interne $\mathcal{T}_{t_i}$, un test binaire Φ_{t_i} conçu pour *optimalement* bipartitionner le sous-ensemble correspondant au nud. Les sous-ensembles résultants alimentent deux nuds enfants $\mathcal{T}_{t_j}$ et $\mathcal{T}_{t_k}$ qui sont ajoutés en tant qu'enfants directs de $\mathcal{T}_{t_i}$. Ce processus est répété jusqu'à ce qu'il n'y ait plus de bipartition possible. Chaque nud feuille contient alors les probabilités de réalisation des étiquettes de la VT à son niveau. Le test binaire de bipartition Φ_{t_i} peut être obtenu en sélectionnant la division du sous-ensemble des données produisant une impureté des étiquettes inférieure à un seuil donné, en se servant du critère de Gini.

Après la construction des arbres de `T`, la classification d'un nouvel échantillon de données x_0 est réalisée en le transmettant à tous les arbres puis en sélectionnant l'étiquette qui est la plus observée sur l'ensemble des feuilles résultats. Concrètement, les nuds internes sont utilisés pour envoyer x_0 à l'un de leurs nuds enfants, en fonction du résultat du test binaire Φ_{t_i} associé, jusqu'à ce qu'il atteigne un nud feuille. Chaque arbre produira une classe locale pour cet échantillon d'entrée x_0 et la classe finale de l'échantillon sera la classe la plus votée dans l'ensemble des arbres. Une illustration visuelle simplifiée du processus de construction et de l'utilisation du classifieur RF est présentée à la figure IV.2.

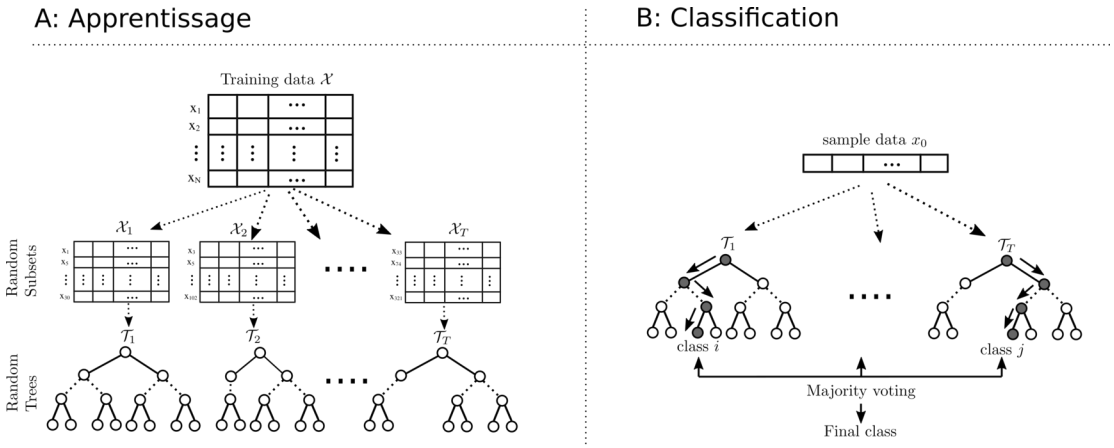


FIGURE IV.2. – Illustration de la classification par forêts aléatoires : un classifieur de forêt aléatoire est un ensemble d'arbres de décision binaires. Chaque arbre est construit à partir d'un sous-ensemble aléatoire des données d'apprentissage. Chaque nud interne dispose d'un test logique permettant de diviser ses données d'entrée en deux sous-ensembles. Les probabilités de prédiction des nuds-feuilles sont utilisées en phase de classification.

Il a été démontré que les forêts aléatoires produisent des performances comparables à SVM dans les problèmes de classification multi-classes (Bosch *et al.*, 2007) tout en maintenant une efficacité de calcul

élevée. La popularité des RF est principalement due à leurs propriétés attrayantes qui incluent : (i) leur efficacité de calcul à la fois en apprentissage et en prédiction, (ii) leur sortie probabiliste, (iii) le traitement transparent d'une grande variété de caractéristiques visuelles (e.g. couleur, texture, forme, profondeur, etc.), et (iv) la robustesse contre les valeurs aberrantes. Par ailleurs, comparativement aux autres techniques de ML pour la classification, les RF offrent un très bon compromis entre la taille de l'ensemble d'apprentissage et la qualité de la classification.

3.1. Sélection de caractéristiques

Souvent, dans un processus d'apprentissage automatique, les caractéristiques décrivant les données ne contribuent pas également à prédire la réponse cible ; dans de nombreuses situations, la plupart des caractéristiques sont en réalité non pertinentes.
De par le principe de leur construction, les arbres de décision effectuent intrinsèquement la sélection des caractéristiques en sélectionnant les points de partage appropriés de données. Étant donné un arbre $\mathcal{T}$, chaque nud interne n_i de $\mathcal{T}$ est affecté à une fonction de test binaire sur une caractéristique f_i qui réalise la meilleure bipartition du sous-échantillon $\mathcal{X}_{\mathcal{T}_i}$ de n_i. La profondeur de n_i dans $\mathcal{T}$ donne l'importance relative de f_i dans la prévisibilité de la variable cible. Les caractéristiques utilisées en haut de l'arbre contribuent à la décision finale de prédiction d'une fraction plus importante des échantillons d'entrée. La fraction attendue des échantillons auxquels ils contribuent peut donc être utilisée pour estimer l'importance relative des caractéristiques. Ainsi, cette information peut être utilisée pour mesurer l'importance de chaque caractéristique de $\mathcal{X}_{\mathcal{T}_i}$. Plus une caractéristique f_i est utilisée dans les points de scission de l'arbre $\mathcal{T}$, plus elle est importante. Cette notion d'importance peut être étendue à des ensembles d'arbres de décision, tels que les forêts aléatoires, en faisant simplement la moyenne de l'importance des caractéristiques de chaque arbre. En faisant la moyenne des estimations de la capacité prédictive sur plusieurs arbres randomisés, la variance d'une telle estimation peut être réduite et utilisée pour la sélection des caractéristiques (Breiman, 2001). C'est ce qu'on appelle la diminution moyenne de l'impureté (Mean Decrease Impurity - MDI) (Louppe, 2014). L'importance d'une caractéristique f_j est calculée en additionnant les diminutions d'impureté Δ pondérées $p(n)\Delta i(s_n, n)$ par p pour tous les nuds n_i où f_j est utilisée, moyenné sur tous les arbres $\mathcal{T}_m$ (pour $m = 1, \ldots, M$) dans la forêt :

$$Imp(f_j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\mathcal{T} \in \mathcal{T}_m} 1(j_n = j)\,[p(n)\Delta i(s_n, n)] \tag{IV.1}$$

Parallèlement, d'autres mesures de l'importance des caractéristiques ont été proposées, telles que la diminution moyenne de la précision (Mean Decrease Accuracy - MDA) (Breiman, 2001, 2002) qui calcule l'erreur de diminution moyenne de la forêt. L'élimination récursive de caractéristiques (RFE) (Granitto *et al.*, 2006 ; Guyon *et al.*, 2002) est une autre approche populaire de la sélection de caractéristiques. Cette approche élimine de manière récursive les caractéristiques à travers les trois étapes suivantes : (1) Apprentissage du classifieur ; (2) calculer le critère de classement pour toutes les caractéristiques ; (3) Supprimer la caractéristique avec le critère de classement le plus petit.
Dans ce chapitre, la sélection des caractéristiques est réalisée via l'approche MDI. En fonction de l'importance individuelle de chaque caractéristique, il est possible de réduire l'ensemble des caractéristiques utilisées pour l'apprentissage du modèle de prédiction automatique. Une valeur de seuil Imp_0 est utilisée

pour élaguer l'arbre de recherche en supprimant les entités dont la valeur d'importance est inférieure à Imp_0. Dans le cas de forêts aléatoires, il est courant d'assigner cette valeur à la moyenne ou à la médiane des valeurs d'importance de toutes les caractéristiques initiales.

3.2. Réglage des hyper-paramètres : Hyperparametrisation

L'objectif ultime d'un algorithme d'apprentissage typique $\mathcal{A}$ est de trouver une fonction f qui minimise certaines pertes attendues $\mathcal{L}(x; f)$ sur des échantillons indépendantes et identiquement distribuées (i.i.d.), x d'une distribution de vérité terrain $\mathcal{G}_x$. Un algorithme d'apprentissage $\mathcal{A}$ est une fonction qui mappe un ensemble de données $\mathcal{X}^{(train)}$ à une fonction f (Bergstra et Bengio, 2012). Cependant, la plupart des algorithmes d'apprentissage ont leurs propres paramètres, appelés hyper-paramètres, qui doivent être définis lors de leur construction. Les hyper-paramètres sont des paramètres qui ne sont pas appris directement par les estimateurs mais qui sont fournis par l'utilisateur lors de la construction de l'estimateur ou qui sont ajustés indirectement aux données. Ils fournissent généralement des informations *à priori* sur la distribution des données à apprendre. Cawley et Talbot (2014) ont constaté qu'il peut être préférable d'adapter les paramètres de l'algorithme directement sur les données (avec un terme de régularisation supplémentaire), plutôt que l'approche habituelle qui les traite comme des hyper-paramètres ajustés via une validation croisée. Il est possible et recommandé de rechercher dans l'espace des hyper-paramètres le meilleur score de validation croisée afin d'optimiser les hyper-paramètres de l'estimateur (Pedregosa *et al.*, 2011). Une procédure de recherche nécessite :

- un estimateur : algorithme à instancier ;
- un espace de paramètre : domaine de valeurs de paramètre ;
- une méthode de recherche ou d'échantillonnage de candidats : comment les valeurs de paramètre sont sélectionnées pour l'instanciation ;
- un schéma de validation croisée : stratégie de réglage des paramètres ; et
- une fonction de score : schéma d'évaluation d'instance.

Les stratégies courantes de recherche dans l'espace des hyper-paramètres comprennent la Recherche par Grille (Grid Search), la Recherche Manuelle (Manual Search) et la Recherche Aléatoire (Randomized Search) (Bergstra et Bengio, 2012). Les deux premières stratégies de recherche procèdent par une instanciation de l'estimateur considéré. Ensuite, une phase d'apprentissage est itérée sur toutes les combinaisons possibles de valeurs de paramètres. La meilleure combinaison est conservée. Randomized Search implémente une recherche aléatoire sur des paramètres, chaque paramètre étant échantillonné à partir d'une distribution sur des valeurs de paramètres possibles. Ceci présente deux avantages principaux, par rapport à une recherche exhaustive : (1) Une limite peut être choisie indépendamment du nombre de paramètres et des valeurs possibles. (2) L'ajout de paramètres qui n'influencent pas les performances ne diminue pas l'efficacité (Pedregosa *et al.*, 2011).

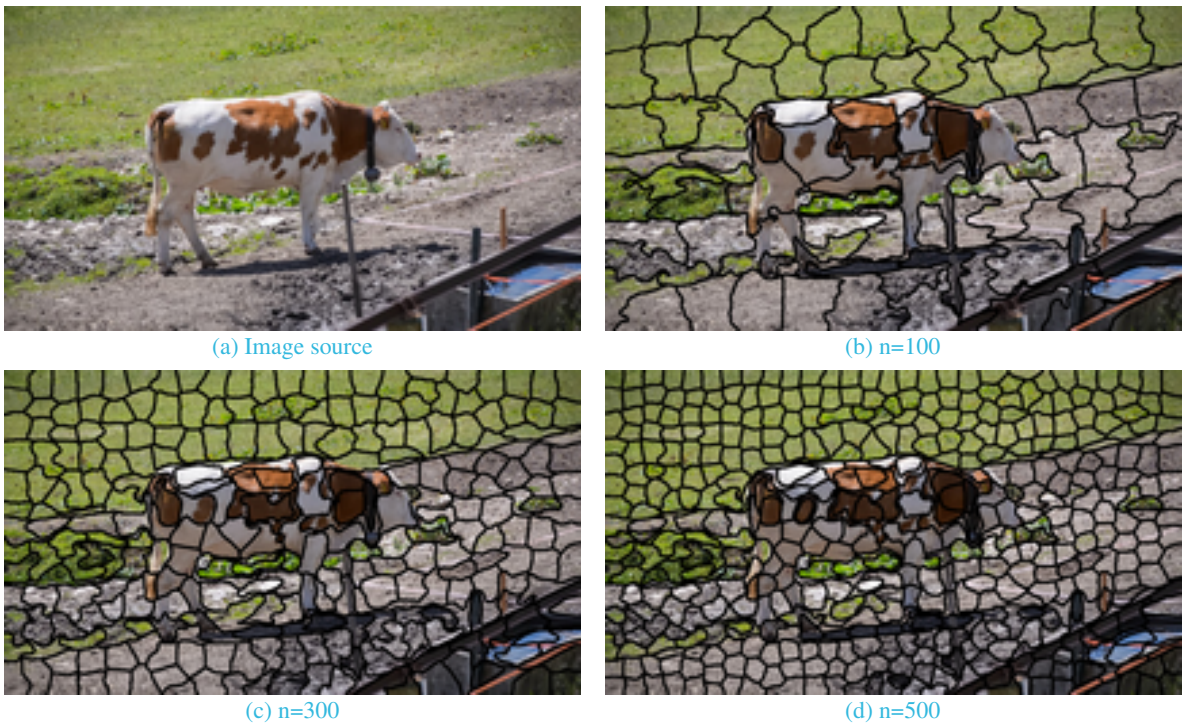
4. Décomposition multi-niveaux d'images en superpixels

Dans cette étape, l'algorithme **CoSLIC** proposé dans le chapitre **III** pour décomposer une image $\mathcal{I}$ en un ensemble de superpixels $\mathcal{S}_p$. Notre algorithme est une extension du SLIC (Achanta *et al.*, 2012) fournissant aux superpixels une meilleur adhérence aux contours de l'image. Le superpixel offre, par rapport au pixel, la possibilité de calculer certaines caractéristiques telles que la texture et la forme. Plusieurs approches

consacrées à cette tâche sont développées dans la littérature (voir chap.**II**, Sect. 3). De la même manière, de nombreuses approches de segmentation basées sur les pixels peuvent être appliquées sur des images sur-segmentées avec des pertes de performances négligeables et un gain de temps considérable.

4.1. Décomposition multi-niveaux

L'observation de la figure IV.3 montre que plus le nombre de superpixels générés est grand, plus la décomposition est fine et meilleure. Ainsi, une simple décomposition en superpixels produira un ensemble de



(a) Image source (b) n=100

(c) n=300 (d) n=500

FIGURE IV.3. – Décomposition d'image en superpixels par l'algorithme CoSLIC. Une image (IV.3a) et sa décomposition correspondante en 100 (IV.3b), 300 (IV.3c) and 500 (IV.3d) superpixels.

superpixels réguliers qui sera utilisé pour l'apprentissage de la fusionnabilité des superpixels. Par conséquent, les données d'apprentissage du classifieur seront composées uniquement de ce type de superpixels i.e. avec des propriétés similaires, telles que la taille, la forme, ... Ceci peut affecter la capacité de prédiction du modèle lorsque des superpixels ayant des propriétés différentes lui seront servis en entrées. Pour pallier ces failles, nous proposons de générer une décomposition en superpixels sur plusieurs niveaux de chaque image d'apprentissage pour constituer une base d'apprentissage du classificateur RF plus diversifiée et donc plus robuste.

En effet, les données d'apprentissage déterminent complètement la portée et l'efficacité du modèle construit. De ce fait au lieu d'utiliser une seule décomposition, des images d'apprentissage, en un nombre fixe α de superpixels, nous générons plusieurs décompositions de chaque image en faisant varier le nombre de superpixels à générer de α_{min} à α_{max} avec une écart prédéfini δ où $\alpha_{min} < \alpha_{max}$ `et` $\delta > 0$. Cela se traduit par une décomposition de l'image à plusieurs niveaux allant de superpixels plus grossiers, pour de petites valeurs de α, à des superpixels plus fins lorsque α augmente comme le montre la figure IV.4. Ainsi, le processus d'apprentissage sera effectué avec différentes vues de résolution de la même image. Cela conduit à une augmentation de la taille (1) des données d'apprentissage et la variété (2) de cas utilisés,

ce qui améliore la robustesse du modèle de classification appris.

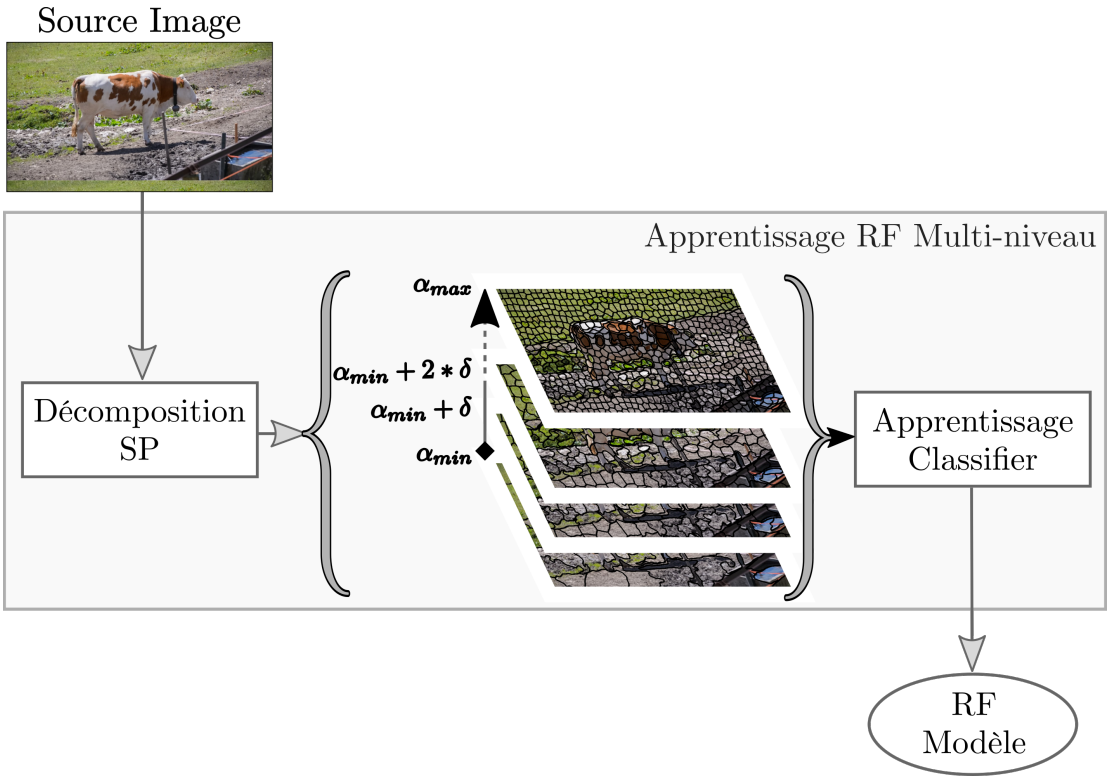


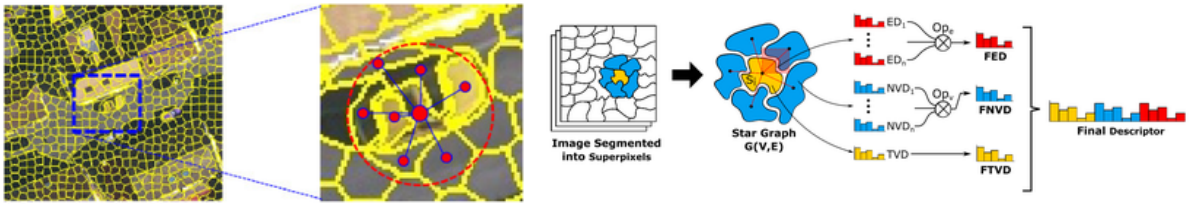
FIGURE IV.4. – Décomposition en superpixels à plusieurs niveaux pour l'apprentissage du classifieur. Étant donné une image source $\mathcal{I}$, plusieurs décompositions en superpixels de $\mathcal{I}$ sont effectuées. Cette décomposition à plusieurs niveaux fournit des échantillons d'apprentissage de différents niveaux de granularité. Ainsi, le modèle appris peut refléter beaucoup plus de cas qu'une décomposition à un seul niveau.

4.2. Contexte local du superpixel

Dans les méthodes de segmentation par croissance de régions, la représentation des régions est une étape cruciale car elle constitue le support de la comparaison des régions et de ce fait celui du processus de croissance. Dans la caractérisation d'une région, l'information relative au contexte local permet une meilleure robustesse face au bruit (Yu et Clausi, 2008). En effet, les zones voisines qui sont ajoutées à la description de la région offrent une vue plus étendue de la région et permettent de confirmer ou d'infirmer la caractérisation intrinsèque à la région.

Les approches de segmentation basées sur les superpixels proposent plusieurs configurations pour caractériser le contexte local d'un superpixel. Parmi les travaux récents, en télédétection Vargas *et al.* (2015) propose d'intégrer le contexte local dans des cartes thématiques représentant les superpixels. Dans ce travail, chaque superpixel de l'image est décrit à l'aide d'un histogramme d'éléments visuels (Visual Words), à l'aide de la méthode de Bag of Visual Words (BoVW). Ensuite, les informations contextuelles sont codées en concaténant la description de superpixel avec une combinaison des histogrammes de ses voisins pour générer un nouveau descripteur de contexte. L'un des principaux inconvénients de cette méthode est le manque de codage explicite des aspects relationnels parmi les caractéristiques extraites des superpixels adjacents. Pour la classification des superpixels en télédétection, Santana *et al.* (fig. IV.5a)

proposent d'exprimer le contexte local d'un superpixel par un vecteur descripteur, qu'ils appellent le descripteur **Star**. Tout d'abord, ils regroupent dans un vecteur les caractéristiques des superpixels voisins par concaténation non ordonnée. Ensuite, ils calculent un vecteur de texture entre le superpixel et chacun de ses voisins à travers un rectangle dont la diagonale est formée par les centres des deux superpixels. Le contexte local correspond à la concaténation du vecteur descripteur du superpixel, des vecteurs des voisins et du vecteur de texture. Ces travaux font une représentation du contexte local par une structure plate contrairement à Audebert *et al.* (2017) où le contexte local est codé par un graphe. A l'inverse de Santana *et al.* ; Vargas *et al.*, les auteurs ici proposent d'utiliser les superpixels qui sont dans un rayon `r` du centre du superpixel courant. L'introduction de `r` permet d'éviter le biais que peut causer les voisins de très petite taille. En plus ils conservent l'ordre spatial des voisins dans la structure de graphe utilisée.



(a) Audebert *et al.* (b) Santana *et al.*

FIGURE IV.5. – Illustration des représentations du contexte local d'un superpixel. IV.5a approche proposée par Audebert *et al.*. IV.5b approche proposée par Santana *et al.*.

La figure IV.6 présente une illustration visuelle de quelques exemples de superpixels accompagnés de leur contextes locaux.

Nous proposons une représentation du contexte local similaire à celle proposé par (Audebert *et al.*, 2017) que nous appliquons au superpixel résultant de la fusion des superpixels candidats. Cependant, pour une paire $p = (a, b)$ voisin, au lieu de considérer les superpixels pris séparément, le calcul du contexte est fait sur la paire p pour lui produire une description unifiée. Cela évite la redondance des voisins communs mais aussi le déséquilibre dans le cas des superpixels présentant une différence importante de taille. Enfin, l'expression du contexte local sur la paire de superpixels permet de s'aligner avec l'idée d'apprentissage basée sur la notion de couple de superpixels, présentée dans les sections suivantes.

En outre, dans (Audebert *et al.*, 2017) le voisinage d'un superpixel a correspond à l'ensemble des superpixels $\{a_i\}$ dans le rayon `r`. Cette approche peut engendrer un déséquilibre lorsqu'un superpixel $a_j \in \{a_i\}$ tel que $|a_j| \gg |a|$ est atteint par `r`. Ainsi, au lieu de considérer un cercle des $\{a_i\}$, nous suggérons de créer un rectangle englobant du voisinage du superpixel dont les mesures dépendent de celles du superpixel courant a. Premièrement, il faut déterminer le rectangle minimum englobant $\mathcal{R}a$ de a, dont nous notons la largeur par `W` et la longueur par `L`. Par la suite, le voisinage à considérer, noté $C(a)$, est contenu dans le rectangle de largeur `2×W`, de longueur `2×L` et centré sur a. Cela nous permet d'éviter la détermination empirique de ces valeurs et garantit l'obtention d'un contexte local auto-adaptatif qui change selon la taille de a. En effet, dans (Audebert *et al.*, 2017) quelque soit la taille du superpixel, `r` reste fixe. Cette contrainte pourrait fausser l'impact du contexte car lorsque celui-ci est trop grand par rapport au superpixel, il dominera la caractérisation du superpixel tandis que son effet pourrait être insensible lorsque

celui-ci est trop petit par rapport au superpixel. Le contexte local ainsi défini est formalisé comme suit :

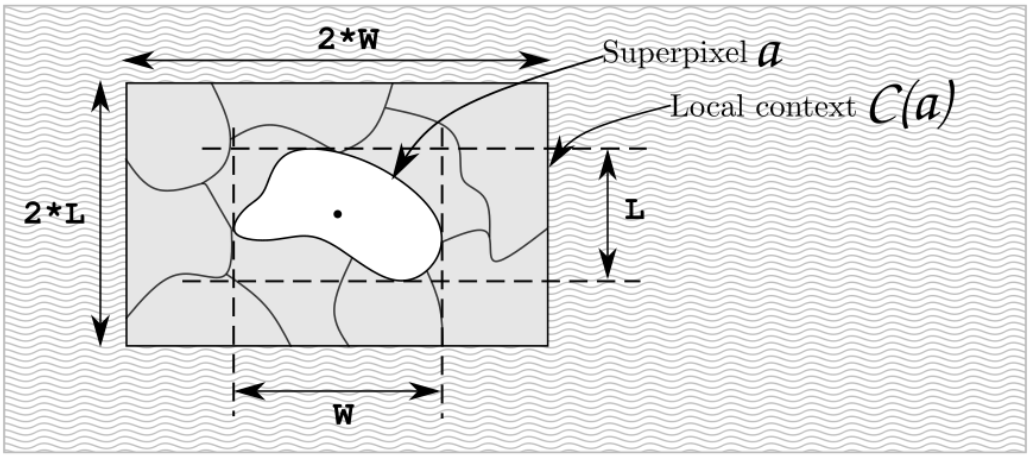
$$C(a) = \Big\{ p_i \mid p_i \in \mathcal{I} \ \texttt{et} \ p_i \notin a \ \texttt{et} \ p_{i_x} \in [x_0 - 2 \times \texttt{W}, x_0 + 2 \times \texttt{W}] \ \texttt{et} \ p_{i_y} \in [y_0 - 2 \times \texttt{L}, y_0 + 2 \times \texttt{L}] \Big\} \tag{IV.2}$$

Avec (x_0, y_0) les coordonnées du centre de a. Le contexte local C est représenté sous forme d'une seule région afin de présenter une vue homogène du voisinage du superpixel.

5. Similarité de superpixels par apprentissage RF

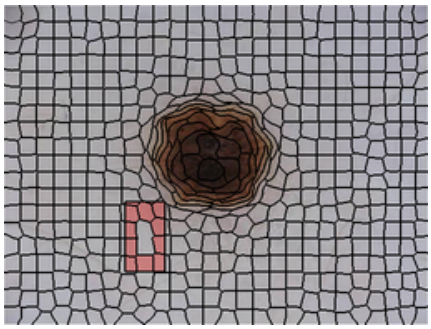
La principale contribution de ce chapitre est axée sur l'apprentissage automatique de la similarité entre deux superpixels donnés en lieu et place d'un calcul par le biais d'une formule mathématique entre les vecteurs descripteurs des superpixels.

Étant donné l'efficacité des classifieurs RF dans les problèmes de classification, nous suggérons de



(a)

(b)



(c)

FIGURE IV.6. – Illustration de descripteur de contexte proposé. IV.6a présente un superpixel a avec son contexte local $C(a)$. IV.6b et IV.6c montrent deux exemples de superpixel et leurs contextes locaux correspondants. Le contexte est vu comme une seule région.

modéliser le problème du calcul de la similarité entre deux régions, dans les approches de segmentation par croissance de régions, par une tâche de classification. À partir de la base d'images initiale $\mathcal{D}$, la phase d'apprentissage est réalisée à l'aide de la base d'apprentissage $\mathcal{D}_{train}$ décrite à la section 3. Pour ce faire, chaque image $\mathcal{I} \in \mathcal{D}_{train}$ est décomposée en un ensemble de superpixels $\mathcal{S}_p$ à travers la méthode de décomposition multi-niveau présentée dans §4.1. Ensuite, $\mathcal{P}$, l'ensemble de toutes les paires de superpixels de $\mathcal{S}_p$ qui est défini par :

$$\mathcal{P} = \{ (a,b),\ \forall\ a,b \in \mathcal{S}_p \text{ and } \mathcal{N}(a,b) = 1\} \tag{IV.3}$$

où la fonction $\mathcal{N}$ exprime l'adjacence spatiale entre deux superpixels, a and b comme suit :

$$\begin{aligned} \mathcal{N} &: \mathcal{S}_p \times \mathcal{S}_p \rightarrow \{0,1\} \\ \mathcal{N}(a,b) &= \begin{cases} 1, & \texttt{si } a \texttt{ est adjacent à } b \\ 0, & \texttt{sinon} \end{cases} \end{aligned} \tag{IV.4}$$

Les éléments de $\mathcal{P}$ peuvent être divisés en deux groupes : les paires *fusionnables* $\mathcal{P}_f$ et les les paires non fusionnables correspondant à $\mathcal{P} \setminus \mathcal{P}_f$. Donc, une paire (a,b) de superpixels adjacents a et b est étiquetée avec `fusionnable` lorsque les deux superpixels appartiennent à la même classe d'objet C_i (cf §2). Inversement, si a et b n'appartiennent pas à la même classe, alors l'étiquette de fusion de (a,b) est `non-fusionnable`. Ainsi, la définition suivante des paires de superpixels fusionnables.

$$\mathcal{P}_f = \{ (a,b),\ \forall\ (a,b) \in \mathcal{P}\ \mid\ \mathcal{N}(a,b) = 1,\ a,b \in C_i \} \tag{IV.5}$$

De là, nous pouvons entrainer un classifieur RF pour apprendre la fusion des paires de superpixels. Après apprentissage, le modèle RF appris sera capable de fournir une probabilité de fusion pour toute paire de superpixels en entrée.

Les données utilisées pour l'apprentissage RF sont constituées par de paires de superpixels accompagnées de leur étiquette de fusion correspondante. Concrètement, chaque superpixel est substitué par son vecteur de descripteurs calculés (cf.§4.2) auxquels le contexte local de la paire est ajouté. Ainsi, considérant deux superpixels a et b avec les vecteurs de caractéristiques respectifs f_a et f_b, le vecteur caractéristique de la paire (a,b) équivaut à la concaténation de f_a, f_b et $C_l(a \cup b)$. L'ensemble des données d'apprentissage $\mathcal{X}$ est donc défini par :

$$\mathcal{X} = \{ (f_a, f_b, C_l(a \cup b), \ell),\ \forall\ (a,b) \in \mathcal{P},\ \forall\ \mathcal{I} \in \mathcal{D} \} \tag{IV.6}$$

où l'étiquette de fusion ℓ de la paire de superpixels (a,b) est donnée par :

$$\ell = \begin{cases} \texttt{fusionnable},\ \texttt{si}\ \ (a,b) \in \mathcal{P}_f \\ \texttt{non} - \texttt{fusionnable},\ \texttt{sinon} \end{cases} \tag{IV.7}$$

Cette phase d'apprentissage du classifieur produit en sortie un modèle de classification des paires de superpixels $\hat{\mathcal{Y}}$ qui est formulé par l'équation IV.8. Ainsi, pour toute paire de superpixels (a,b), ce modèle donne l'étiquette prédite de la paire (a,b).

$$\hat{\mathcal{Y}}(a,b) = y,\ y \in \{ \texttt{fusionnable}, \texttt{non-fusionnable} \} \tag{IV.8}$$

En fait, le modèle de classificateur RF fournit, pour la paire donnée, une valeur de probabilité ω pour chaque étiquette. Nous utilisons donc la forme donnant la sortie probabiliste de l'équation IV.8 comme suit :

$$\begin{cases} \tilde{\mathcal{Y}}(a,b) = \{\omega_f, \omega_n\}, \\ 0 \leqslant \omega_f, \omega_n \leqslant 1 \ \texttt{et} \ \ \omega_f + \omega_n = 1 \end{cases} \tag{IV.9}$$

L'étiquette attribuée est celle avec la probabilité la plus élevée.

6. Segmentation par agrégations itératives

Rappelons que notre objectif est la segmentation d'une image par croissance de régions dans laquelle la similarité des régions est estimée par un classifieur RF. Plus précisément, la phase de segmentation commence par une décomposition en superpixels de l'image d'entrée $\mathcal{I}$. Deuxièmement, l'ensemble des superpixels obtenus $\mathcal{S}_p$ est utilisé pour construire un graphe d'adjacence de région $\mathcal{G}$. Ensuite, une probabilité d'agrégation est calculée entre chaque paire de nuds adjacents dans le graphe en utilisant le classifieur RF entraîné précédemment et finalement selon leur probabilité de fusion calculées, les paires de nuds adjacents sont successivement fusionnés. Cette deuxième étape est répétée de manière itérative jusqu'à atteindre la convergence du résultat. La phase de segmentation est effectuée sur les images de $\mathcal{D}_{test}$ et la probabilité d'agrégation d'une paire (a,b) est donnée par sa probabilité ω_f d'être étiquetée comme `fusionnable`. L'algorithme IV.1 résume ce processus.

Algorithme IV.1 : Segmentation

Entrées : $\mathcal{I}$: image à segmenter, $\hat{\mathcal{Y}}$:modèle de prédiction de similarité
Output : $\mathcal{G}$:graphe des régions finales

```
1 S_p := CoSLIC(I);
2 G := GAR(S_p);
3 tant que (convergence(G) = Faux) faire
4     G := EtiqueterAretes(G, Ŷ);
5     G := FusionnerNoeuds(G);
```

6.1. Modélisation de l'image par le graphe d'adjacence de régions (GAR)

L'approche de segmentation proposée utilise une structure de graphe indirect pondéré pour modéliser les relations spatiales de l'ensemble des superpixels de l'image. Initialement, à partir de l'ensemble des superpixels précédemment générés, un graphe d'adjacence des régions (GAR) est créé. La figure IV.7 présente quelques exemples des graphes générés. Le GAR d'une image $\mathcal{I}$, noté $\mathcal{G} = (V, E)$, est entièrement défini par l'ensemble des nuds V et l'ensemble des arêtes E représentant toutes les paires de superpixels liés par une relation de voisinage. Le graphe est formalisé comme suit :

$$\mathcal{G} = \{ (v,e) \mid v \in V \ \texttt{et} \ e \in E \} \tag{IV.10}$$

avec $V = \mathcal{S}_p$ et $E = \{ e \mid \forall e, \exists (a,b) \in {\mathcal{S}_p}^2 \mid e = (a,b) \ \texttt{et} \ \mathcal{N}(a,b) = 1 \}$. Chaque nud de $\mathcal{G}$ représente un superpixel et se caractérise par un vecteur de 3×107 caractéristiques de couleur qui inclut, mais sans s'y limiter à, un histogramme de couleur de 26 cases et quelques valeurs statistiques sur les quatre premiers

moments de couleur centrale, y compris la moyenne (E), l'écart type (σ), l'asymétrie ($Skew$) et le kurtosis ($Kurt$), calculé comme suit :

$$E(a) = \frac{1}{N}\sum_{i}^{N}(a_i) \tag{IV.11}$$

où a_i est le i^{th} pixel du superpixel a et N est le nombre total de superpixels in a. Les trois autres valeurs sont calculées sur la base de la valeur de E comme suit $\sigma(a) = \sqrt{m_2(a)}$, $Skew(a) = \frac{m_3(a)}{[\sigma(a)]^3}$ and $Kurt(a) = \frac{m_4(a)}{[\sigma(a)]^4}$ avec

$$m_j(a) = \frac{1}{N}\sum_{i}^{N}[a_i - E(a)]^j \tag{IV.12}$$

Ensuite, pour toute arête du GAR e, reliant deux nuds a et b, le modèle $\tilde{\mathcal{Y}}$ est appliqué à la paire (a, b) pour estimer sa probabilité de fusion ω_f ; celle-ci correspondant au poids de l'arête e.

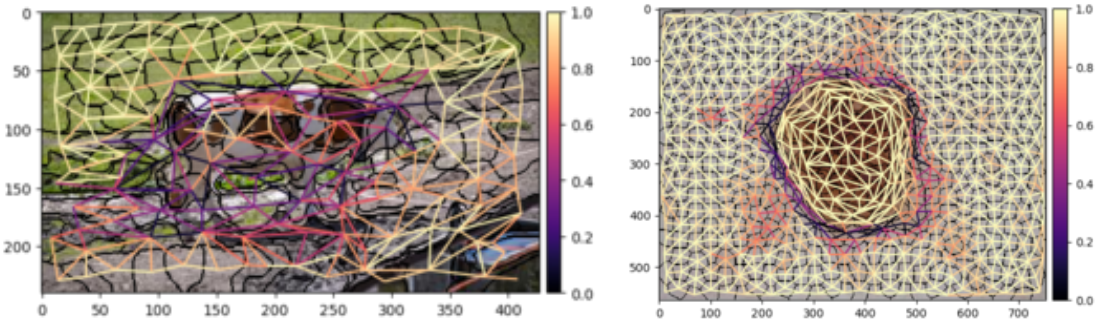


FIGURE IV.7. – Quelques exemples illustratifs de la représentation d'images par graphe de superpixels. Dans ces exemples le nombre de superpixels est réduit (150 pour IV.7a et 500 pour IV.7b) afin d'obtenir des graphes visuellement simples. La couleur des arêtes reflète la valeur de son poids. plus le poids est élevé plus l'arête est sombre.

6.2. Agrégations itératives de superpixels basée sur le GAR

La structure de graphe utilisée permet de résoudre le problème de segmentation en utilisant des algorithmes de coupe de graphe. Par conséquent, la segmentation de l'image peut être interprétée comme un problème de partition graphe. La littérature de la théorie des graphes présente plusieurs travaux qui proposent différentes approches pour résoudre ce problème (Sangsefidi *et al.*, 2017 ; Silva, 2017 ; Zhang et He, 2017).

Dans notre cas nous proposons une approche simple mais efficace qui se compose de deux étapes exécutées de manière récursive jusqu'à la convergence. Tout d'abord, étant donné le GAR pondéré $\mathcal{G}$ de l'image d'entrée, l'arête de poids maximum du graphe est sélectionnée comme étant la meilleure arête ε candidate à la fusion exprimée par

$$\varepsilon = \text{argmax}_{e \in E}[f(e)] \tag{IV.13}$$

La fonction f assigne à chaque arête e son poids ω_e. Ainsi, l'arête ε donne les deux nuds voisins les plus susceptibles d'être fusionnés dans tous les nuds du graphe $\mathcal{G}$. Alors, en supposant que ε est formée par la paire (a, b), les deux nuds a et b sont fusionnés en un nouveau nud c qui les remplace dans le GAR $\mathcal{G}$. En conséquence c sera lié à tous les nuds voisins de a et b. Chaque nouvelle arête e', entre c et un nud voisin n,

est pondérée par un nouveau poids ω' calculé comme suit :

$$\omega' = \omega_f \ \mid \tilde{\mathcal{Y}}(c,n) = \{\omega_f, \omega_n\} \tag{IV.14}$$

À l'issue de cette opération un nouveau GAR $\mathcal{G}' = (V', E')$ représentant la même image $\mathcal{I}$ est obtenu, dans lequel

- $V' = \{V \setminus \{a,b\}\} \cup \{c\}$
- $E' = \{E \setminus \{ E_a, E_b \}\} \cup \{ E_c \}$ avec

$E_a = \{ e \mid e = (a,n); \forall n \in V \texttt{ et } e \in E\}$ est l'ensemble de toutes les arêtes impliquant le nud a, $E_b = \{ e \mid e = (b,n); \forall n \in V \texttt{ et } e \in E\}$ correspond à l'ensemble de toutes les arêtes impliquant le nud b et $E_c = \{ e \mid e = (c,n); \forall n \in V \texttt{ et } (a,n) \in E \texttt{ ou } (b,n) \in V\}$ représente l'ensemble de toutes les arêtes impliquant le nud nouvellement ajouté c.

La sélection de la meilleure paire de nuds à fusionner et la mise à jour du GAR sont répétées de manière itérative sur le nouveau GAR jusqu'à ce que tous les poids des arêtes dans le graphe soient inférieurs ou égaux à une valeur seuil ω_0.

6.3. Raffinement de la sélection d'arête

Une bonne approche de segmentation produit des résultats dans lesquels l'inertie interne des régions est maximale et celle entre ces régions est minimale. Les approches itératives majoritairement intègrent ces concepts par une fonction objective $\mathcal{J}$ qui est formée de plusieurs termes associés aux objectifs ciblés. A chaque itération, la fonction choisit les meilleures régions à fusionner. A partir du graphe $\mathcal{G}$ de l'image, nous proposons une fonction $\mathcal{J}_E$ d'énergie à minimiser afin de mieux sélectionner l'arête de $\mathcal{G}$ à chaque itération. Par conséquent, étant donné le graphe $\mathcal{G}$, la fonction de sélection de la meilleure arête à fusionner est mise à jour comme suit :

$$\varepsilon = \text{argmax}_{e \in E} \left[\mathcal{J}_E(e) \times f(e)\right] \tag{IV.15}$$

$\mathcal{J}_E$ se compose de deux parties : un coefficient de taille $\jmath_s$ et un coefficient de voisinage $\jmath_c$.

$$\mathcal{J}_E(a,b) = \jmath_s(a,b) + \jmath_c(a,b) \tag{IV.16}$$

Le coefficient de taille $\jmath_s$ permet une pondération de la similarité entre la paire de superpixels (a,b) candidats à la fusion en fonction de leur tailles.

$$\jmath_s(a,b) = \frac{1}{2} \times \left[S_\Delta^{x_0,y_0} \left(\frac{min(|a|,|b|)}{max(|a|,|b|)} \right) + \frac{|a \cup b|}{|\mathcal{I}|} \right] \tag{IV.17}$$

où $|a|$ calcule la taille du superpixel a et $S_\Delta^{x_0,y_0}(x) = \frac{1}{\pi} \arctan\left[\frac{x-x_0}{\Delta}\right] + y_0$ est la fonction de normalisation $arctan$ centrée sur x_0, de pente Δ et décalée de y_0. Il faut noter que $y_0 = \texttt{0.5}$ correspond à la valeur qui produit une normalisation des valeurs entre $\texttt{0}$ et $\texttt{1}$. La figure IV.8 montre quelques cas particuliers de la normalisation en utilisant la fonction $S_\Delta^{x_0,y_0}$. $|a \cup b|$ correspond à la région couvrant les superpixels a et b. La première composante de $\jmath_s$ assure le regroupement des superpixels de tailles proches tandis que la seconde composante permet de favoriser la fusion de superpixels de petites tailles.

Quant au coefficient de voisinage $\jmath_c$, il assure que les deux nuds extrêmes de l'arête choisie possèdent un voisinage similaire. D'abord, nous définissons par contexte d'un nud a, le vecteur $u = \{u_i\}_{i \in [0,..,2]}$ formé

par les trois premiers moments de l'ensemble des poids des arêtes incidentes à a. Ensuite, pour une arête $e = (a, b)$, le contexte $\jmath_c$ de e est défini par :

$$\jmath_c(a,b) = \begin{cases} \alpha = min_i \left\{ S_\Delta^{x_0,-y_0} \left(\frac{|u_i - v_i|}{u_i + v_i} \right) \right\} \ si\ \alpha \geqslant 0 \\ 0\ sinon \end{cases} \tag{IV.18}$$

u et v représentent les vecteurs contextes respectifs de a et b. $y_0 \in \mathbb{R}$ est, dans ce cas, un paramètre utilisé pour réguler la similarité des distributions de contextes des nuds. A titre d'exemple $y_0 =$ `0.85` permet d'imposer une similarité d'au moins `85%` entre les nuds, en termes de poids d'arêtes incidentes respectives.

7. Évaluation expérimentale

Dans toutes les expériences, le nombre d'arbres dans une forêt est défini à `50` et la profondeur d'arbre est fixée à `10` niveaux. $\alpha_{min}, \alpha_{max}$ et δ sont respectivement fixés à `200, 850` et `200`. Sauf indication contraire, le nombre de superpixels initiaux à la phase de segmentation est égal à `500`. Dans ce travail, la valeur de ω_0 est définie à `0.55` afin de permettre uniquement la fusion entre des nuds voisins avec une probabilité supérieure à `55%`. L'évaluation de l'approche proposée est divisée en deux parties : l'évaluation de la qualités des prédictions du modèle appris et celle de la segmentation produite.

7.1. Bases d'images d'évaluation

L'approche de segmentation proposée est validée sur deux bases d'images : DAVIS (Pont-Tuset *et al.*, 2017) et ISIC (International Skin Imaging Collaboration) 2018.

DAVIS 2017 est une base de séquences vidéo publiquement disponible et conçue pour la segmentation d'objets dans les séquences vidéo. Elle rassemble une collection de vidéos disponibles sous forme d'images. Ainsi, les images de la même vidéo fournissent un jeu d'images à scène unique. La vérité terrain d'une

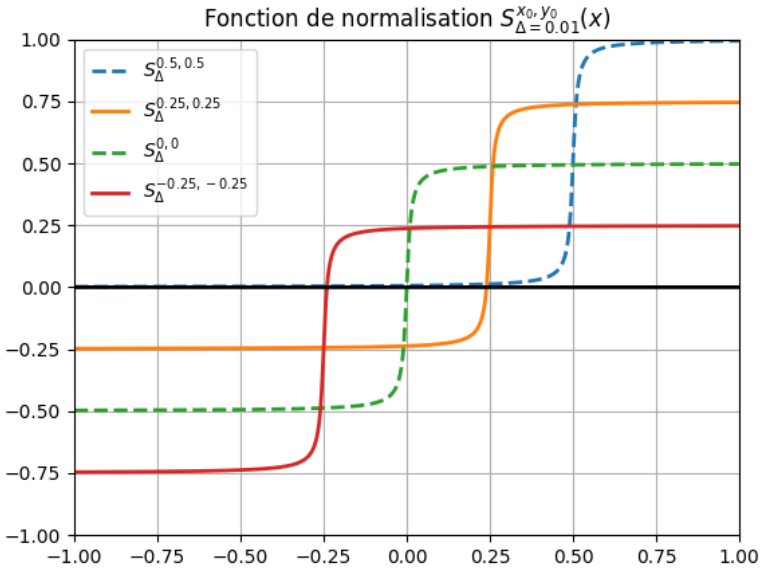


FIGURE IV.8. – Fonction de normalisation $S^{x_0,y_0}_{\Delta=0.01}(x) = \frac{1}{\pi} \arctan\left[\frac{x-x_0}{\Delta}\right] + y_0$. Nous avons utilisé les valeurs de $(x_0 = 0.5, y_0 = 0.5)$ pour la normalisation de $\jmath_s$ et $(x_0 = 0.5, y_0 = -0.9)$ pour celle de $\jmath_c$.

séquence est formée par des masques des frames qui séparent l'objet d'avant-plan et le reste du frame qui est considéré comme arrière-plan. Dans les séquences d'apprentissage, chaque frame est fourni avec une VT correspondante alors que dans les séquences de test seuls le premier possède une annotation VT. La base fournit `60` séquences vidéo pour les phases d'apprentissage. Quelques exemples d'images de la base sont présentés dans la figure IV.9 [1].

Pour chaque séquence vidéo, `n = 2` images sélectionnées aléatoirement avec leurs annotations sont fournies au classifieur pour l'apprentissage. Nous utilisons les séquences d'apprentissage dans toutes nos expérimentations. Le classifieur entrainé est ensuite utilisé pour segmenter le reste des frames de la même séquence vidéo. L'évaluation est effectuée par comparaison avec la vérité terrain des frames fournie.

FIGURE IV.9. – Quelques exemples d'images de la base DAVIS 2017 et leur VT associées [2]

La seconde base d'images utilisée pour nos expériences est celle fournie par ISIC (International Skin Imaging Collaboration). ISIC est un partenariat entre les universités et l'industrie visant à faciliter l'application de l'imagerie numérique de la peau afin d'étudier et d'aider à réduire la mortalité par mélanome. Les archives ISIC contiennent plus de `13000` images de lésions cutanées étiquetées comme étant bénignes ou malignes. Nous avons utilisé les images préparées pour la compétition ISIC18 pour la tâche de segmentation [3]. Il s'agit d'environ `2600` images qui sont fournies chacune avec le masque VT des lésions associé. La figure IV.10 présente quelques exemples des images de la base ISIC 18.
À cause de la différence des conditions de collecte des images, nous avons constitué un sous dossier contenant `100` images qui présentent des conditions similaires. Par la suite, nous prenons `n = 2` images aléatoirement avec leurs annotations pour l'apprentissage du classifieur. La phase de segmentation est réalisée sur le reste des images.

1. Image prise de https://davischallenge.org/davis2017/code.html(visité le 30/09/2017)
3. https://challenge2018.isic-archive.com/ visité le 25/11/2018

7.2. Évaluation de l'approche proposée

Dans cette section, nous présentons les résultats obtenus par l'approche proposée sur les bases d'évaluation présentées ci-haut. Les performances de l'approche sont évaluées, dans un premier temps dans les sections 7.2.1 et 7.2.2, par composante. Ensuite, nous proposons des comparaisons des performances de notre approche avec celles de quelques approches de la littérature dans les sections d'après.

7.2.1. Similarité de superpixels

Cette section présente l'évaluation de la qualité de la mesure de similarité proposée dans ce chapitre afin d'évaluer la qualité des prédictions du modèle proposé. En considérant l'ensemble des paires de superpixels issus de la décomposition d'une image, nous notons par :

- VP (Vrais Positifs) : le nombre de paires de superpixels positifs et classifiés comme tels.
- FP (Faux Positifs) : le nombre de paires de superpixels négatifs et classifiés comme positifs.
- FN (Faux Négatifs) : le nombre paires de superpixels positifs et classifiés comme négatifs.
- VN (Vrais Négatifs) : le nombre de paires de superpixels négatifs et classifiés comme tels.

Conséquemment, les quatre métriques suivantes sont calculées sur les graphes résultats, comme suit :

- **Sensibilité** : mesure la capacité à donner un résultat positif lorsqu'une hypothèse est vérifiée.

$$\textbf{Sensibilité} = \frac{VP}{VP + FN} \tag{IV.19}$$

- **Spécificité** : mesure la capacité à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée.

$$\textbf{Spécificité} = \frac{VN}{VN + FP} \tag{IV.20}$$

- **MSE** : mesure la moyenne des carrés des erreurs, c'est-à-dire la différence moyenne au carré entre les valeurs estimées et ce qui est estimé.

$$\textbf{MSE} = \frac{FP + FN}{FP + FN + VP + VN} \tag{IV.21}$$

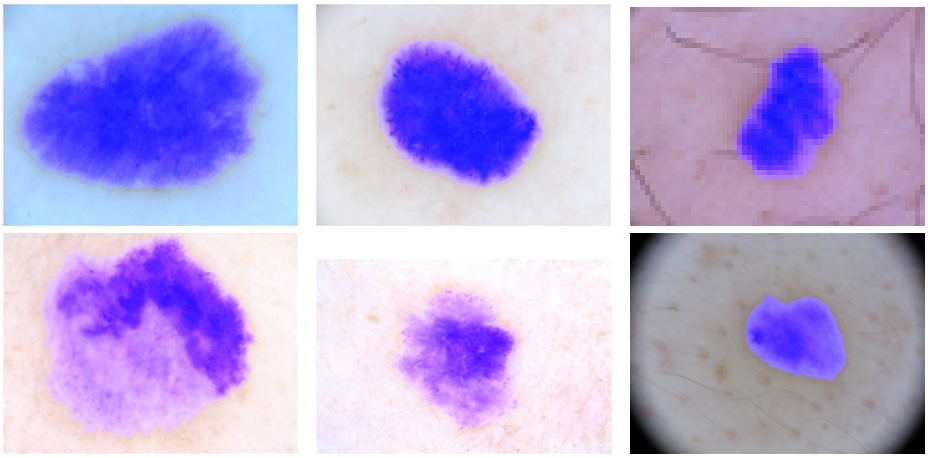


FIGURE IV.10. – Quelques exemples d'images de la base ISIC 2018 et leur VT associées.

- **F-mesure** : mesure de la précision du modèle de prédiction.

$$\textbf{F-mesure} = \frac{2 \times VP}{2 \times VP + FN + FP} \tag{IV.22}$$

Pour ce faire, pour chaque probabilité de fusion ω de paire de superpixels, nous étiquetons par 1 lorsque $\omega > 0.5$, sinon la prédiction est marquée comme incorrecte en lui assignant l'étiquette 0.
Nous commençons par l'évaluation de la mesure de similarité par le modèle de classification appris. Les résultats des expériences sont résumés dans le tableau IV.1 et les figures IV.11 et IV.12 en présentent une synthèse graphique.

TABLE IV.1. – Résultats d'évaluation du calcul de similarité des superpixels en utilisant la décomposition simple (RF-SDC), multi-niveaux (RF-MDC) et les caractéristiques de contexte (RF-MDC+CTX). Les meilleurs résultats sont en **gras**

		Sensitivity ↑			Specificity ↑			MSE ↓			F-mesure ↑		
		Min	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**
DAVIS 2017	**RF-SDC**	0.743	0.896	0.863	0.273	0.865	0.523	0.146	0.215	0.169	0.803	0.876	0.857
	RF-MDC	0.886	0.945	0.927	0.419	0.802	0.662	0.087	0.147	0.102	0.872	0.929	0.918
	RF-MDC+CTX	**0.952**	**0.991**	**0.977**	**0.560**	**0.898**	**0.783**	**0.031**	**0.065**	**0.043**	**0.948**	**0.982**	**0.974**
ISIC 2018	**RF-SDC**	0.783	0.874	0.852	**0.139**	0.484	0.217	0.155	0.271	0.186	0.795	0.871	0.853
	RF-MDC	0.858	0.925	0.907	0.122	0.468	0.241	0.105	0.200	0.132	0.859	0.921	0.905
	RF-MDC+CTX	**0.913**	**0.982**	**0.962**	0.094	**0.575**	**0.259**	**0.054**	**0.138**	**0.078**	**0.917**	**0.972**	**0.958**

Décomposition multi-niveaux en superpixels

La décomposition en superpixels est une étape fondamentale du travail présenté dans ce chapitre. Nous avons proposé deux approches pour utiliser la décomposition du superpixel afin de fournir des données d'apprentissage : une décomposition simple et une décomposition à plusieurs niveaux (cf.§4.1). Dans les figures IV.11 et IV.12 nous présentons les courbes d'évaluation qui comparent la qualité de prédiction du modèle obtenu à partir d'une simple décomposition en superpixels et celle à plusieurs niveaux.

La décomposition multi-niveaux a été ajoutée pour permettre au modèle appris d'appréhender des superpixels de tailles différentes de ceux initiaux. Comme le montre les résultats, dans le cas de la décomposition simple, le nombre de superpixels étant fixé à $n = 500$, la qualité de la similarité est la meilleur sensiblement autour de cette valeur. Alors que pour des décompositions multi-niveaux en superpixels, les courbes de la qualité de similarité entre superpixels présente une allure strictement croissante jusqu'à $n \geqslant 750$. Les propriétés des superpixels dans l'approche à un seul niveau ne correspondent pas à celles des superpixels initiaux, ce qui explique la constance de la qualité des résultats. Dans ce cas, l'approche multi-niveaux permet une adaptation des superpixels grâce aux superpixels de niveaux supérieurs avec des propriétés similaires et produit ainsi de meilleurs résultats.
Étant donné l'observation de meilleurs résultats de la décomposition en plusieurs niveaux par rapport à une décomposition simple, dans la suite des expériences la décomposition en superpixels est faite sur plusieurs niveaux.

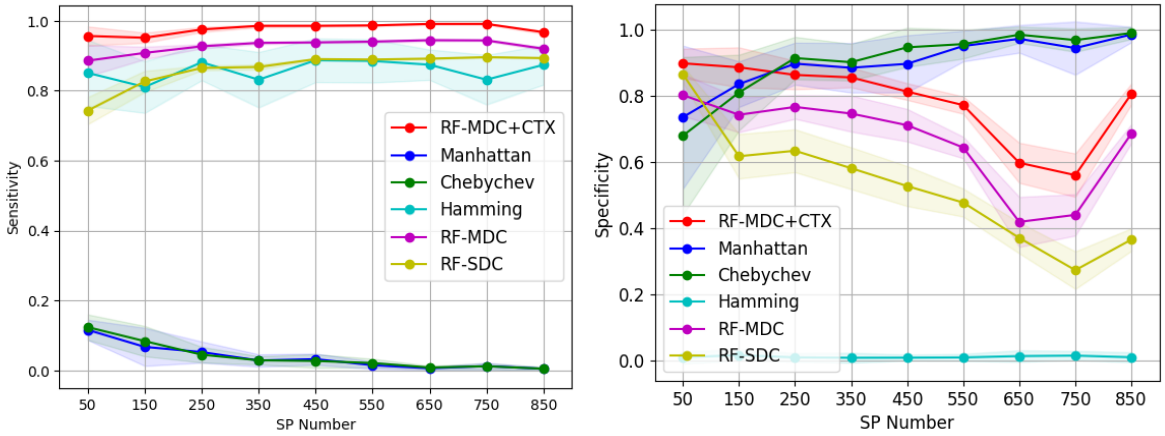
Effets du contexte local

La décomposition multi-niveaux en superpixels permet de corriger la dépendance du modèle vis-à-vis du nombre de superpixels initiaux mais n'introduit aucune relation entre les paires de superpixels appris. Nous avons ainsi introduit le voisinage du superpixel dans sa description dans les phases d'apprentissage du modèle RF afin de tenir compte de cet aspect. Les figures IV.11 et IV.12 montrent les résultats sur la qualité du modèle appris sur des descripteurs sans le contexte local et lorsque celui-ci est considéré.

L'analyse des courbes présentées révèle une amélioration claire de la qualité des résultats. L'ajout des descripteurs de contexte (RF-MDC+CTX) a permis un gain net dans tous les critères et sur toutes les bases d'évaluation de la similarité prédite par le modèle RF appris.

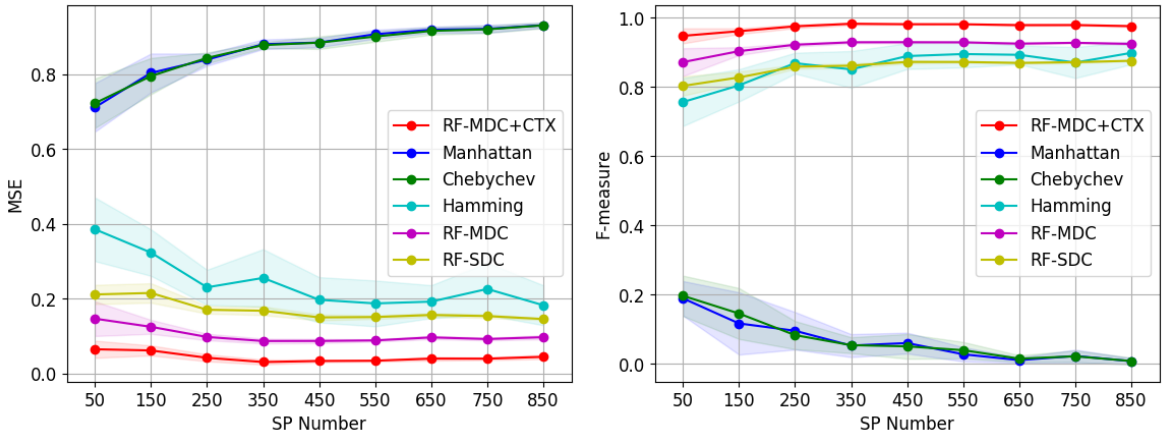
Mesures de similarité classiques

Cette section présente une comparaison de la qualité de la probabilité de fusion des superpixels générée par le classifieur appris avec trois (3) mesures de similarité bien connues pour calculer les poids des GAR



(a) Courbe de la sensibilité en fonction du nombre de superpixels.

(b) Courbe de la spécificté en fonction du nombre de superpixels.



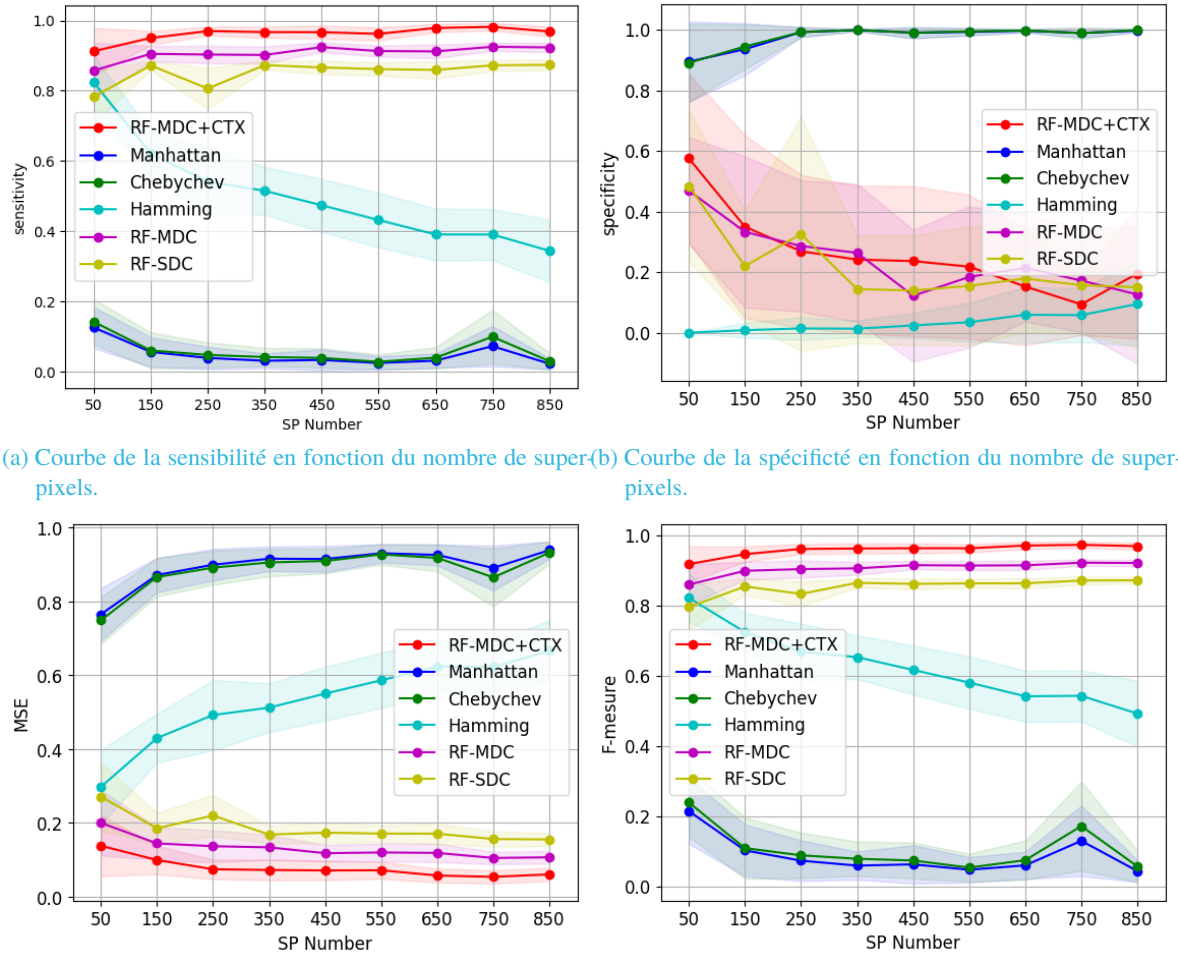
(c) Courbe MSE en fonction du nombre de superpixels.

(d) Courbe F-mesure en fonction du nombre de superpixels.

FIGURE IV.11. – Comparaison de la mesure de similarité du modèle pour une décomposition en superpixels simple (RF-SDC), multi-niveau (RF-MDC), multi-niveau avec descripteurs de contexte (RF-MDC+CTX) et les distances classiques en fonction du nombre de superpixels initiaux sur la base DAVIS 2017.

d'une image : Hamming (HAM) (Hamming, 1950), City-block (CTB) et la mesure de Chebyshev (CHEB). A partir du GAR de l'image, chacune de ces mesures est utilisée pour étiqueter le GAR. Les poids des arêtes correspondent à la valeur de similarité calculée sur les vecteurs caractéristiques superpixels des nuds de l'arête correspondante. Le tableau IV.2 résume l'évaluation des résultats comparatifs de la méthode de calcul de similarité proposée et des trois mesures sélectionnées. Les résultats visuels sont fournis dans les figures IV.11 et IV.12.

Les résultats montrent que la mesure de Hamming produit les plus mauvais résultats dans les deux jeux de données et sur tous les critères d'évaluation. Les mesures City-block et Chebyshev produisent des résultats assez similaires. L'approche proposée surpasse ces mesures de similarité sur tous les critères d'évaluation, à l'exception de Specificty. En effet, notre approche permet de doubler la qualité dans les trois critères, par exemple en termes de sensibilité le meilleur résultat de notre approche de `0,962` est deux fois mieux que le meilleur des mesures de comparaison qui est de `0,504`. Il en va de même pour la MSE et la F-mesure, respectivement pour `0,078` contre `0,531` et `0,958` contre `0,627`.



(a) Courbe de la sensibilité en fonction du nombre de superpixels.

(b) Courbe de la spécificté en fonction du nombre de superpixels.

(c) Courbe MSE en fonction du nombre de superpixels.

(d) Courbe F-mesure en fonction du nombre de superpixels.

FIGURE IV.12. – Comparaison de la mesure de similarité du modèle pour une décomposition en superpixels simple (RF-SDC), multi-niveau (RF-MDC), multi-niveau avec descripteurs de contexte (RF-MDC+CTX) et les distances classiques en fonction du nombre de superpixels initiaux sur la base ISIC 18.

TABLE IV.2. – Résultats d'évaluation du calcul de similarité Superpixel pour la mesure de Hamming, City-block, Chebyshev et notre approche en utilisant la décomposition multi niveau et les caractéristiques de contexte (RF-MDC+CTX). Les meilleurs résultats sont en **gras**.

		Sensitivity ↑			Specificity ↑			MSE ↓			F-mesure ↑		
		Min	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**
DAVIS 2017	**Hamming**	0.811	0.887	0.859	0.008	0.016	0.011	0.183	0.386	0.242	0.756	0.899	0.859
	City-block	0.004	0.115	0.037	**0.735**	0.985	0.900	0.712	0.932	0.867	0.008	0.190	0.065
	Chebyshev	0.004	0.123	0.039	0.679	**0.990**	**0.905**	0.723	0.931	0.866	0.009	0.197	0.069
	RF-MDC+CTX	**0.952**	**0.991**	**0.977**	0.560	0.898	0.783	**0.031**	**0.065**	**0.043**	**0.948**	**0.982**	**0.974**
ISIC 2018	**Hamming**	0.344	0.824	0.504	0.000	0.095	0.034	0.298	0.666	0.531	0.492	0.821	0.627
	City-block	*0.023*	*0.126*	*0.049*	**0.895**	**0.999**	0.976	0.765	0.939	0.894	0.045	0.215	0.089
	Chebyshev	0.028	0.142	0.059	0.891	**0.999**	**0.977**	0.749	0.932	0.885	0.054	0.239	0.106
	RF-MDC+CTX	**0.913**	**0.982**	**0.962**	0.094	0.575	0.259	**0.054**	**0.138**	**0.078**	**0.917**	**0.972**	**0.958**

Cela montre bien que le modèle RF appris gère mieux le calcul de similarité entre paires de superpixels que les distances classiques. Ceci confirme l'hypothèse de la supériorité du principe d'apprentissage du modèle puisque nous considérons les paires de superpixels comme des entités élémentaires au lieu de superpixels pour calculer la similarité. De plus, dans le processus d'apprentissage, des superpixels visuellement différents peuvent être considérés comme similaires s'ils appartiennent au même objet sémantique. Les mesures de similarité classiques ne disposent d'aucun mécanisme leur permettant de gérer ce comportement. L'approche proposée pour la mesure de la similarité a pour principal inconvénient de ne pas détecter les superpixels de l'arrière-plan et du premier plan comme appartenant à différents objets sémantiques, comme le montrent les résultats de la spécificité.

La figure IV.13 présente des résultats visuels comparatifs des mesures de similarité. La couleur des arêtes dans le GAR indique la valeur de similarité entre les superpixels de nud. Plus l'arête est claire, plus la valeur est élevée. Comme on peut le voir sur la figure, le modèle RF-MDC+CTX proposé produit des résultats plausibles et sépare distinctement le premier plan de l'arrière-plan comme supposé. La mesure de Hamming révèle certaines incohérences dans les poids des arêtes, en particulier au voisinage de la limite avant-plan/arrière-plan. Ce comportement entraîne des résultats de segmentation incorrects. De plus, le modèle proposé produit de bonnes valeurs de similarité pour les superpixels à l'intérieur de l'objet du premier plan, même pour ceux ayant des caractéristiques visuelles distinctes. Cela montre la principale force de l'approche de similarité par apprentissage proposée par rapport aux mesures calculées.

7.2.2. Segmentation basée sur graphe

À partir du graphe valué $\mathcal{G}$ de l'image, la segmentation est réalisée par regroupements successifs de pairs de superpixels voisins par ordre décroissant du poids de l'arête qui les relie. La qualité de la segmentation est quant à elle évaluée par les quatre critères de segmentation suivants, qui sont calculés sur les résultats finaux des expériences.

- Probabilistic Rand Index (**PRI**) (Unnikrishnan *et al.*, 2007) mesure la proportion de pixels de la

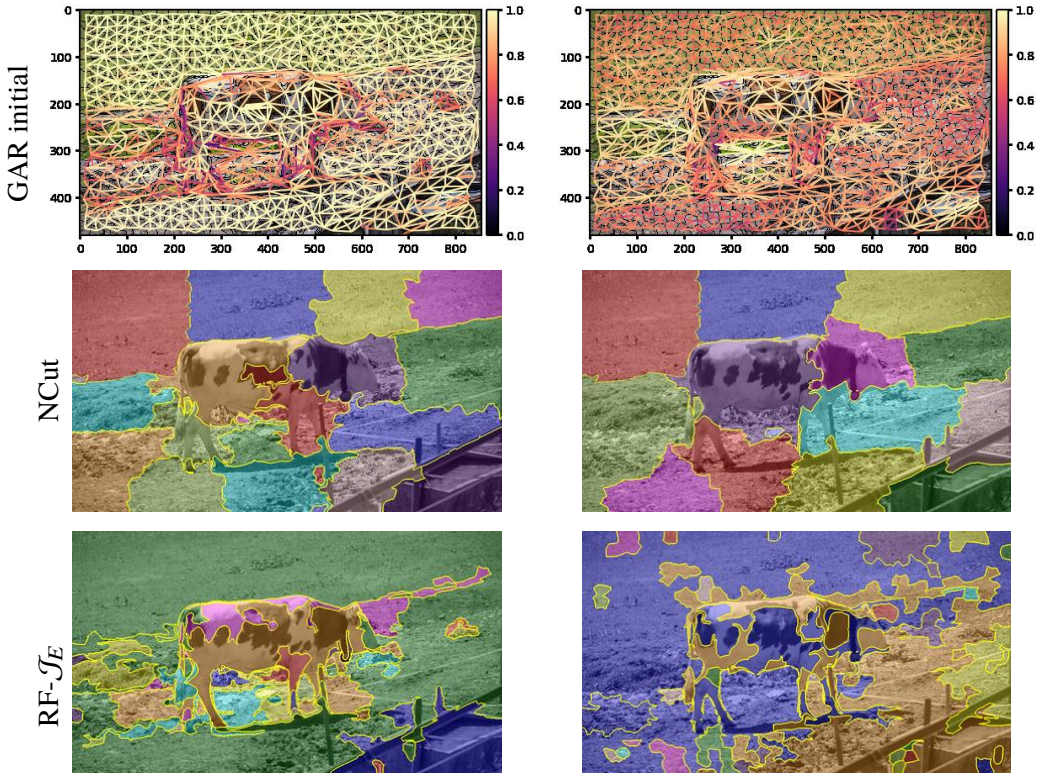


FIGURE IV.13. – Illustration de la mesure de similarité. Chaque colonne donne les résultats du GAR initial créé pour une mesure de similarité. Les résultats de notre modèle RF-MDC+CTX sont présentés dans la première colonne. La mesure de Hamming est présentée dans la deuxième colonne. La première ligne donne le GAR pondéré de l'image. La deuxième et la troisième rangée montrent respectivement l'image segmentée en utilisant NCut et notre approche. Les régions finales sont superposées sur l'image source.

segmentation qui portent les mêmes étiquettes que la segmentation de la base.

- Variation of Information (**VoI**) (Meilă, 2007) mesure la quantité de pixels regroupés de manière aléatoire dans la segmentation, sans aucun indice dans la vérité terrain.
- Boundary Displacement Error (**BDE**) (Freixenet *et al.*, 2002) mesure l'erreur de déplacement moyenne d'un pixel de contours et les pixels contours les plus proches dans la segmentation VT.
- Global Consistency Error (**GCE**) (Martin *et al.*, 2001) mesure la mesure dans laquelle une segmentation peut être considérée comme un raffinement de l'autre. Les segmentations liées de cette manière sont considérées comme cohérentes, car elles peuvent représenter la même image segmentée à différentes échelles.

Plus la valeur PRI est élevée, meilleure est la segmentation, contrairement à VoI, BDE et GCE dont les valeurs élevées dénotent une qualité de segmentation faible.

Raffinement de sélection d'arête

Nous avons proposé un coefficient de pondération du poids des arêtes afin de réguler les regroupements pour aboutir à un résultat de segmentation cohérent. Ce coefficient est composé de deux termes

prépondérants. En fonction du composant de pondération choisi, nous présentons les résultats de la segmentation sur les différentes bases d'images dans la figure IV.14 et le tableau IV.3.

TABLE IV.3. – Comparaison des résultats de la segmentation. Évaluation des résultats de la segmentation à l'aide de la NCut, de TCut et des agrégations de superpixels proposées pour la segmentation (RF-J_S, RF-J_C, RF-$\mathcal{J}_E$). Les meilleurs résultats sont en **gras**.

		PRI ↑			VoI ↓			BDE ↓			GCE ↓		
		Min	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**
DAVIS 2017	**RF-J_S**	**0.372**	0.539	**0.457**	**1.676**	**2.802**	**2.248**	**11.459**	**16.888**	**14.300**	**0.003**	0.157	0.053
	RF-J_C	0.353	**0.545**	0.450	1.771	2.990	2.330	12.392	17.075	14.636	**0.003**	0.138	0.046
	RF-$\mathcal{J}_E$	0.354	0.541	0.442	1.868	2.979	2.415	12.241	17.052	14.751	**0.003**	**0.137**	**0.043**
ISIC 2018	**RF-J_S**	**0.255**	**0.969**	**0.812**	0.270	**3.649**	**1.306**	1.105	174.925	21.385	**0.000**	0.313	0.071
	RF-J_C	0.249	**0.969**	0.803	**0.269**	3.747	1.367	1.105	174.925	21.221	**0.000**	0.313	0.068
	RF-$\mathcal{J}_E$	0.252	**0.969**	0.799	0.268	3.712	1.420	**1.101**	**40.050**	**13.009**	0.016	**0.242**	**0.064**

L'observation des résultats comparatifs entre les différents coefficients de pondération de la fonction sélection de la meilleure arête montre deux orientations principales. Premièrement, le coefficient de pondération par la taille des superpixels (J_S) affiche les meilleurs résultats sur les deux bases d'évaluation en termes de PRI et VoI. Cette tendance confirme l'effet supposé de ce coefficient qui consiste à équilibrer la similarité entre des superpixels par le ratio de leur taille respective. Deuxièmement, les meilleurs résultats selon les critères d'évaluation BDE et GCE sont donnés par le coefficient final ($\mathcal{J}_E$). En effet, ce dernier ajoute au $jmath_s$ une pondération par l'entourage du superpixel à travers le $jmath_c$. Cela, comme le montre les résultats, améliore la qualité de la sélection en augmentant la cohérence des superpixels (GCE) et en réduisant les erreurs de contours (BDE).

Comparaison avec d'autres approches de segmentation

La génération d'une segmentation d'image à partir d'un GAR pondéré de l'image est une autre contribution principale de ce travail. Nous évaluons cette contribution en comparant notre approche avec trois autres techniques populaires de segmentation à base de graphe appliquées sur le GAR généré : la coupure normalisée (NCut) (Shi et Malik, 2000) et la coupure par seuil (TCut). L'approche NCut mesure le coût de la bipartition du graphe en tant que fraction des connexions totales aux arêtes de tous les nuds du graphe. La segmentation est réalisée en calculant de manière récursive une bipartition optimale du graphe jusqu'à la stabilité du résultat. L'algorithme de seuil TCut fusionne simplement les nuds de manière récursive jusqu'à ce qu'il n'y ait plus d'arêtes de poids supérieur à une valeur de seuil donnée. Les résultats sont résumés dans le tableau IV.4.

Les figures IV.15 et IV.16 exposent quelques résultats visuels de la segmentation finale pour le NCut, le TCut et l'approche d'agrégation proposée. Pour la plupart des images, notre technique tend à sur-segmenter l'image source. NCut et TCut génèrent des résultats d'image sous-segmentés par rapport à la VT fournie. Les résultats de l'évaluation de la technique de segmentation sont présentés dans le tableau IV.4. Bien que

la méthode Threshold Cut obtienne les meilleurs résultats par rapport à la plupart des critères utilisés, la technique proposée offre le meilleur résultat BDE et surpasse le NCut en termes de GCE. Cette performance témoigne d'une homogénéité sous-optimale de région et qui peut être expliquée par le manque de vision globale de notre approche sur le graphe lors des agrégations. En effet, seuls les nuds directement impliqués dans la fusion sont considérés pour la re-estimation du poids des nouvelles arêtes.

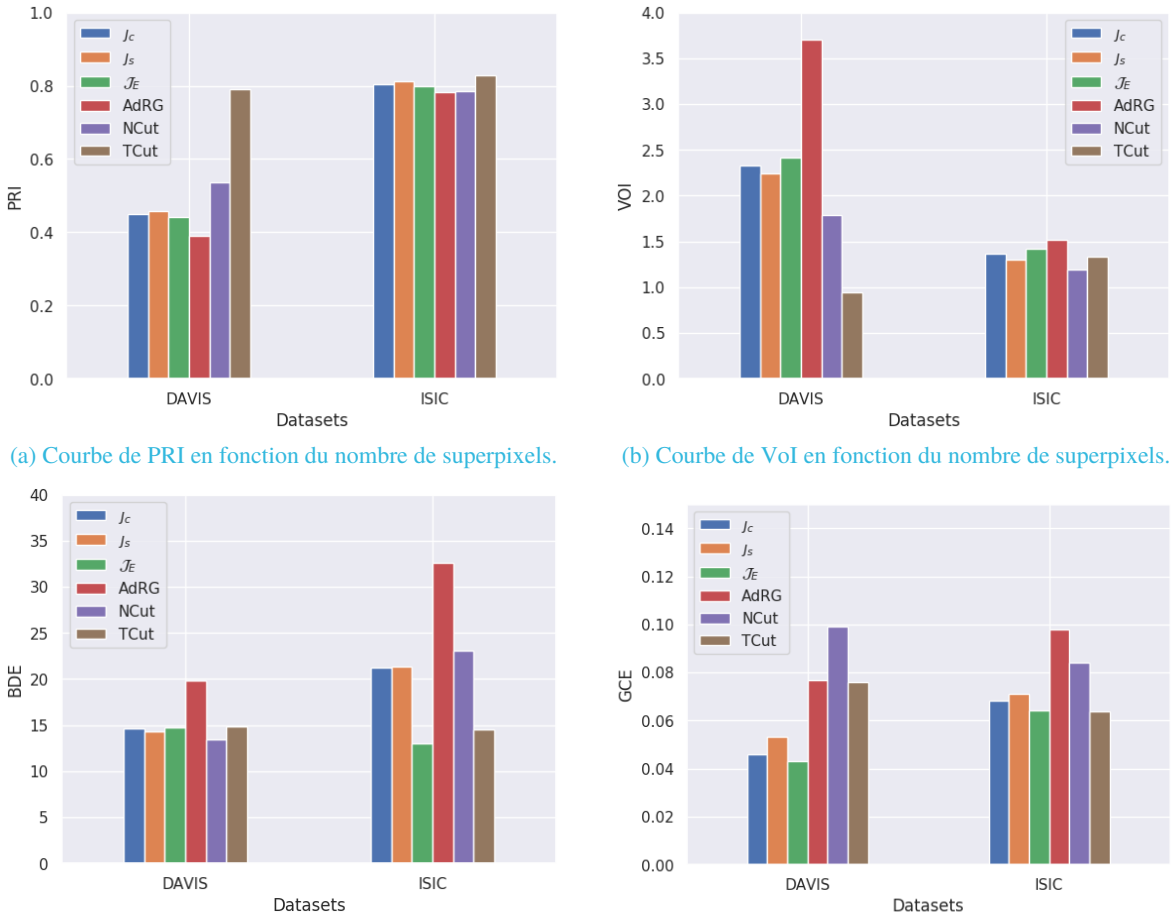
Comparée par rapport à l'approche (**AdRG**) proposée dans le chapitre III, la stratégie d'intégration des connaissances exposée dans ce chapitre obtient les meilleurs résultats de segmentation, sur tous les critères d'évaluation et dans les deux jeux de données.

8. Conclusions et perspectives

Ce chapitre propose une stratégie d'intégration des connaissances dans une approche de segmentation à travers l'apprentissage automatique. Il présente deux contributions principales :

1. un mécanisme basé sur l'apprentissage artificiel pour la fusion de superpixels et
2. une technique d'agrégation basée sur des graphes pour réaliser la segmentation d'images.

Nous avons proposé une approche de fusion de superpixels intégrée dans une approche de segmentation



(a) Courbe de PRI en fonction du nombre de superpixels.

(b) Courbe de VoI en fonction du nombre de superpixels.

(c) Courbe BDE en fonction du nombre de superpixels.

(d) Courbe GCE en fonction du nombre de superpixels.

FIGURE IV.14. – Comparaison des performances du coefficient de sélection meilleure arête pour la segmentation en fonction du nombre de superpixels.

TABLE IV.4. – Résultats comparatifs de la segmentation. Évaluation des résultats de la segmentation en utilisant le Normalized Cut (NCut), le Threshold Cut (TCut), l'agrégation adaptative (AdRG) et l'approche proposée par sélection de meilleure arête (RF-$\mathcal{J}_E$). Les meilleurs résultats sont en **gras**.

		PRI ↑			VoI ↓			BDE ↓			GCE ↓		
		Min	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**
DAVIS 2017	**NCut**	0.379	0.879	0.535	0.654	2.569	1.793	9.531	17.361	**13.442**	0.024	0.201	0.099
	TCut	**0.549**	**0.944**	**0.790**	**0.470**	**1.642**	**0.945**	**5.893**	27.600	14.805	0.006	0.172	0.076
	AdRG	0.342	0.575	0.389	2.622	4.445	3.708	17.231	21.175	19.831	0.044	0.155	0.077
	RF-$\mathcal{J}_E$	0.354	0.541	0.442	1.868	2.979	2.415	12.241	17.052	14.751	**0.003**	**0.137**	**0.043**
ISIC 2018	**NCut**	0.291	0.969	0.785	**0.269**	**3.026**	**1.198**	1.389	174.925	23.088	**0.000**	0.292	0.084
	TCut	**0.361**	**0.973**	0.830	0.337	4.131	1.339	2.192	45.783	14.498	0.018	0.145	**0.064**
	AdRG	0.203	0.962	0.782	0.588	6.032	1.515	1.466	80.093	32.611	0.015	**0.212**	0.098
	RF-$\mathcal{J}_E$	0.252	0.969	0.799	0.268	3.712	1.420	**1.101**	**40.050**	**13.009**	0.016	0.242	**0.064**

d'images basée sur l'agrégation de régions. Cette approche utilise un classifieur par apprentissage artificiel pour apprendre la similarité des couples de superpixels au lieu d'utiliser des mesures de similarité explicites entre superpixels. En particulier, nous avons utilisé le classifieur forêts aléatoires (RF) pour prédire la probabilité de fusion entre toute paire de superpixels voisins. Cela a permis de gérer l'écart sémantique dans l'estimation de la similarité et de manipuler des objets avec des parties hétérogènes. Une approche de décomposition multi-niveaux d'images est utilisée pour offrir au classifieur une meilleure capacité d'apprentissage. En outre l'introduction d'une description contextuelle des superpixels dans leur caractérisation permet d'avoir un champ caractéristique plus vaste de ces derniers pour obtenir un apprentissage plus cohérent des similarités.

D'autre part, étant donné que l'image est représentée par une structure de graphe d'adjacence de région pondérée par des probabilités de fusion, la segmentation est réalisée par un algorithme d'agrégation des nuds du graphe. Cet algorithme est basé sur la sélection de la meilleure paire de superpixels fusionnant à chaque itération en utilisant une fonction de similarité objective. Les résultats obtenus sur les jeux de données DAVIS 2017 et ISIC 2018 indiquent une amélioration substantielle de la performance par rapport aux approches de l'état-de-l'art, à la fois visuellement et quantitativement.

Cependant, équilibrer le ratio du nombre de superpixels de premier plan/arrière-plan en phase d'apprentissage peut permettre une meilleure prédiction de la similarité. En outre, une représentation hiérarchique de la décomposition à plusieurs niveaux en superpixels peut améliorer les performances du classifieur et permettre une intégration aisée et plus robuste du contexte local grâce aux caractéristiques hiérarchiques.

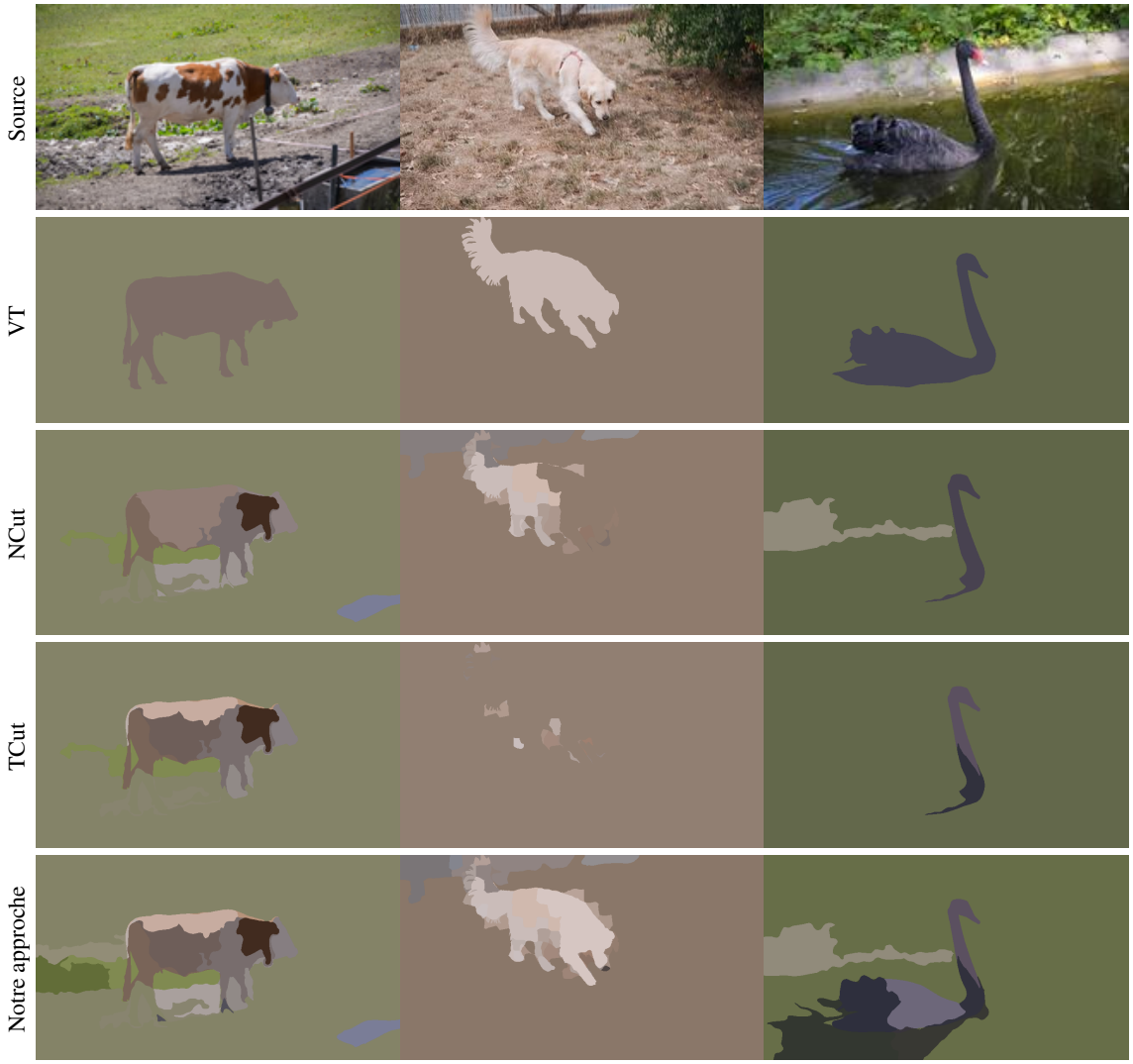


FIGURE IV.15. – Comparaison de la segmentation par coupure de graphe sur la base DAVIS 17. Les images sources sont données dans la première ligne, suivie de la vérité terrain dans la deuxième. Les troisième, quatrième et cinquième lignes donnent respectivement les résultats de segmentation pour le NCut, le TCut et l'approche proposée.

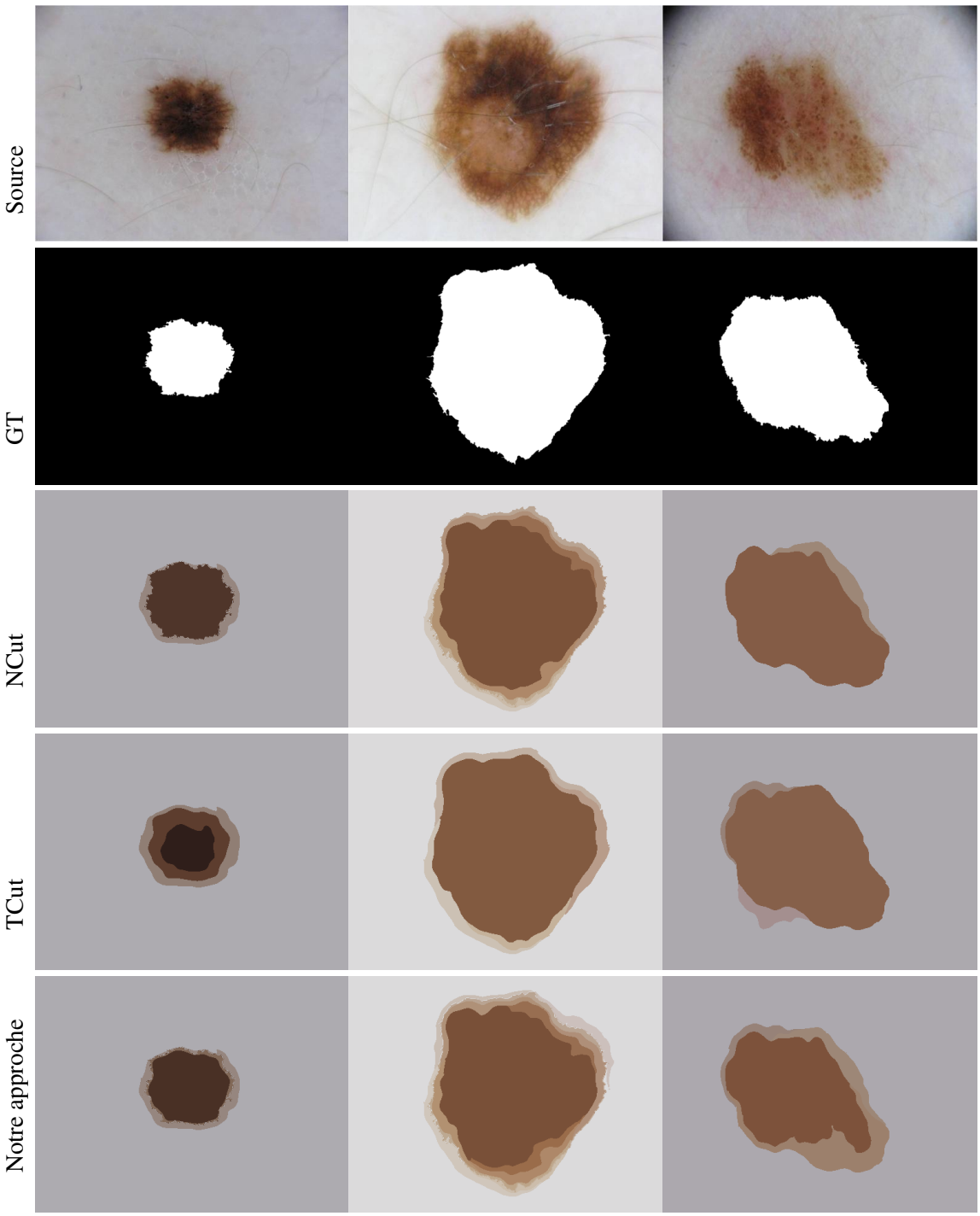


FIGURE IV.16. – Comparaison de la segmentation par coupure de graphe sur la base ISIC 18. Les images sources sont données dans la première ligne, suivie de la vérité terrain dans la deuxième. Les troisième, quatrième et cinquième lignes donnent respectivement les résultats de segmentation pour le NCut, le TCut et l'approche proposée.

Conclusions

Ce chapitre évalue le travail présenté dans ce manuscrit et dégage des conclusions générales. Il présente également des travaux futurs et donne des idées et des réflexions sur certaines perspectives.

> *Everyone is a genius.* ***But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid****.*
>
> *Albert E.*

L'interprétation d'images est une thématique cruciale aujourd'hui au regard des besoins, des technologies et des données disponibles. Ce travail de thèse aborde une étape primordiale de l'interprétation qui est la segmentation d'images. Cette étape constitue le socle de l'interprétation puisqu'elle est responsable de la génération des régions élémentaires qui aboutiront aux objets de la scène interprétée.

Dans cet objectif, nous avons proposé une approche de segmentation qui procède par décomposition en régions élémentaires, appelées superpixels qui seront classifiés et regroupés à travers une étape de raisonnement sur les connaissances qui les caractérisent. Nous avons ainsi dégagé trois zones d'intérêt dans le cycle de segmentation qui ont constitué nos principales contributions développées chacune dans un chapitre du manuscrit.

— Segmentation par propagation de connaissance au lieu de segmentation par regroupement

— Similarité multi-niveau et comparaison adaptative au lieu de critère rigide

— Similarité par apprentissage automatique au lieu de calcul par formule classique.

Afin de situer convenablement ces contributions, nous avons préalablement exposé les différentes composantes d'un système d'interprétation d'images dans le premier chapitre de ce manuscrit.

1. Segmentation dans l'interprétation

Le chapitre « **Segmentation pour l'interprétation de scène** » a présenté un état-de-l'art des approches et des systèmes d'interprétation des scènes avec une attention particulière portée sur les connaissances et leur intégration dans ces approches. Nous avons d'abord présenté une définition et un positionnement des connaissances dans les approches d'interprétation suivies de leur utilité ainsi que les techniques permettant leur utilisation dans la phase de segmentation. Ensuite, nous avons dressé un inventaire non exhaustif des systèmes d'interprétation de scène mettant particulièrement l'accent sur les connaissances utilisées de même que leur intégration. Le chapitre s'est termine par une analyse des méthodes de segmentation de type croissance de régions, qui sont intégrables dans le processus d'interprétation.

Une approche d'interprétation de scènes sous-entend la définition du problème de vision de la scène à interpréter et la manière avec laquelle ce problème est résolu. Plusieurs approches ont été proposées dans la littérature. Chaque approche présente une solution sous son angle de vue, appelé sa *philosophie*. Mais, elles discutent principalement des connaissances impliquées dans l'interprétation et du raisonnement qui permet d'aboutir à celle-ci. Quant aux systèmes, ils constituent une implémentation de l'approche d'interprétation. Les problèmes invoqués aux niveau de cette implémentation tournent au tour des formalismes pour représenter les connaissances, leur organisation architecturale. La partie du raisonnement est généralement déterminée par le formalisme choisi. Ainsi, nous avons vu que tout système d'analyse automatique de données qui se veut efficace doit intégrer des connaissances relatives aux données qu'il traite. Cette intégration peut prendre plusieurs formes selon les propriétés des connaissances impliquées. Les connaissances sont caractérisées par leur type, leur formalisme de représentation et leur niveau de description.

Entre l'approche théorique et l'implémentation en système de l'interprétation d'image, la tâche de segmentation se retrouve imprécisément définie puisque dans les approches elle n'est pas explicitement adressée. Ainsi, les systèmes doivent la définir en fonction de leurs priorités intrinsèques. Dans un objectif de généricité, nous avons proposé de considérer une approche de segmentation de type croissance de régions

car l'interprétation est par nature un processus itératif et progressif. D'ailleurs, l'algorithme schématique des approches de segmentation par croissance de régions présente trois principale étapes : l'initialisation, le calcul de similarité et le regroupement des régions. Conséquemment, nous avons exposé les travaux de l'état-de-l'art dans chacune de ces étapes afin d'offrir une meilleure appréciation de nos contributions.

2. Segmentation par propagation de connaissances

L'approche globale de segmentation d'image proposée dans ce travail définit un modèle d'interprétation de scène qui simule le comportement de l'expert humain face à une tâche d'interprétation de scène. Cette idée étant grandement inspirée du travail d'un radiologue lors d'une séance d'interprétation d'images radiologiques, par conséquent le modèle proposé est fortement influencé par le comportement du radiologue à la tâche. En étudiant le comportement de l'expert humain il apparait clairement que son modèle d'appréhension de la tâche de segmentation n'appartient à aucune des grandes famille des approches de segmentation classiques parmi lesquelles les approches par contours et les approches régions. Il procède plutôt par reconnaissance des zones facilement identifiables. Ensuite, il "*estime*" la nature du reste de l'image par propagation des connaissances en partant de ces zones déjà reconnues. De cette analyse, nous avons retenu et modélisé trois aspects importants du processus d'interprétation d'image :

1. Le processus de segmentation n'est pas nécessairement séquentiel comme la plus part des techniques de segmentations qu'on rencontre, mais plutôt une suite de décisions pouvant remettre en cause leurs prédécesseurs. L'essentiel étant à la fin d'avoir la meilleure classification des régions. L'interprétation ne doit pas être limitée par la segmentation.
2. Le processus de caractérisation d'une zone d'intérêt n'est pas strictement monotone i.e. que l'expert peut aller d'une vue centrée sur la zone à vue plus large incluant ses voisines pour ensuite retourner vers la vue contenant uniquement la zone et vis-versa.
3. Lors de la décision plusieurs sources d'informations sont sollicitées et fusionnées pour une meilleure certitude.

Par conséquent, l'approche proposée simule ce comportement en se basant sur les informations du domaine des images, et se repartit sur quatre principales étapes qui sont la *focalisation* d'attention, la *propagation* des connaissances et la *fusion* d'information. Sachant que les approches de segmentation à base de connaissances se composent essentiellement de (1) l'acquisition des connaissances et du (2) raisonnement sur ces connaissances. Nous avons distingué deux catégories de connaissances à intégrer dans l'approche. Les connaissances *a priori* qui sont fournies par l'expert du domaine et les contraintes visuelles qui sont extraites automatiquement par l'approche. L'essentiel des connaissances *a priori* consiste en la caractérisation de l'ensemble des objets rencontrés dans le type d'image d'étude, aussi appelé *classes thématiques*. Cet ensemble est aussi appelé ensemble de classes modèles et est noté par $C = \{C_1, C_2, \ldots, C_K\}$. Chaque classe est décrite par descripteurs de bas niveau (numériques) et de haut niveau (thématiques) $\mathcal{F}_{C_k} = (H_k, \mu_k, T_k, Pr_k)$.

L'ensemble de ces connaissances est représenté par une structure compacte de graphe d'adjacence dans laquelle chaque nud contient une classe du paradigme orienté objet. Les nuds représentent les classes modèles accompagnées par leurs caractéristiques tandis que le graphe reflète leurs relations spatiales dans une image typique du domaine d'application.
Les contraintes visuelles sur les images sont modélisées par la silhouette $\mathcal{S}_{il}$ et les contours globaux $\mathcal{G}_C$.

La première permet de capter l'organisation schématique des objets dans l'image. La seconde contrainte visuelle modélise l'information complémentaire à la silhouette. En effet, les contours globaux représentent les bordures (frontières) qui viennent définir les limites précises des objets. Les contours sont utilisés pour ajuster les similarités entre des régions voisines.

La phase de raisonnement regroupe toutes les étapes permettant de produire une image segmentée à partir d'une image en entrée en utilisant les connaissances précédemment collectées. Il s'agit donc dans cette étape d'utiliser les connaissances pour manipuler les opérateurs du traitement d'images tels que le filtrage, la classification ou la fusion des régions pour aboutir à la segmentation de l'image en entrée. Nous proposons une phase de raisonnement itérative qui se compose d'une étape de *focalisation d'attention*, suivie d'une étape de la *propagation et l'intégration des connaissances* pour finir par une étape de *fusion des régions* similaires.
L'étape de focalisation permet de sélectionner, parmi les superpixels initiaux, ceux qui possèdent des connaissances suffisantes permettant de les caractériser avec la moindre erreur. Ces superpixels sont appelés les germes. À partir des germes, l'étape de propagation utilise les connaissance disponibles au niveau des germes afin de déterminer la nature du reste des superpixels, jusqu'ici non qualifiés. Une étape de fusion permet de regrouper les superpixels assignés aux mêmes classes thématiques pour former les régions de l'image segmentée.

D'abord, nous avons proposé un raisonnement probabiliste qui utilise les atouts de cette théorie, par rapport aux représentation purement numériques, pour aboutir à la segmentation de l'image. Les résultats de l'évaluation permettent de constater le gain obtenu. Ensuite, nous avons remplacé les distribution des probabilités par celles de possibilités afin de gérer l'imprécision de connaissances. Cet ajout a permis de mieux séparer les distributions des classes initiales et ainsi permettre une meilleure classification des superpixels. Les résultats comparatifs présentent une validation de cette hypothèse.

3. Segmentation par croissance adaptative

La croissance de régions est une technique de segmentation populaire qui consiste à fusionner des régions avec les pixels voisins similaires, de manière itérative. Nous avons utilisé les superpixels comme régions initiales au lieu des pixels. Cela offre plusieurs avantages parmi lesquels leur efficacité de calcul, la réduction de la taille du problème et la possibilité de calcul des caractéristiques de haut niveau (Machairas *et al.*, 2016). Cependant, cette adoption des superpixels pose deux nouveaux problèmes : la mesure de la similarité entre les superpixels et les dépendances des superpixels. La mesure de similarité fait référence à la quantification de la similarité entre deux superpixels. Néanmoins, un algorithme de croissance de régions dépend intrinsèquement de l'ordre de traitement des pixels de l'image (Mehnert et Jackway, 1997). Tant tôt à chaque fois que plusieurs pixels sont à la même distance de leurs pixels voisins ou lorsqu'un pixel est à la même distance de plusieurs régions voisines.

Partant de là, dans le chapitre « **Croissance des régions adaptative** » nous avons proposé une approche de segmentation d'image basée sur une agrégation itérative de régions qui présente trois contributions principales.

Tout d'abord, une extension pour l'algorithme SLIC est proposée afin de fournir une meilleure adhérence aux contours. Deuxièmement, nous avons proposé une mesure de similarité robuste entre les régions qui

utilise à la fois des critères globaux multi-échelles et une comparaison locale à la frontière commune. Elle intègre aussi des informations sur les frontières pour éviter la fusion d'objets superposés. Troisièmement, une stratégie de fusion est conçue pour contrôler les agrégations de régions au fur des itérations en utilisant un seuil de similarité adaptatif. Cette stratégie garantit que les agrégations se produisent dans l'ordre de similarité décroissante. L'efficacité des contributions proposées a été évaluée par des comparaisons avec des approches de segmentation bien établies sur la base d'image BSDS500.

4. Similarité par apprentissage automatique

Dans le chapitre IV nous avons abordé la problématique du fossé sémantique dans le calcul de la similarité inter-région. Dans le processus de segmentation par croissance de régions, la similarité entre les entités élémentaires est un point crucial. À l'inverse des techniques classiques qui calculent la similarité par application de distance aux descripteurs extraits, nous avons proposé dans le chapitre « **Similarité des superpixels par apprentissage** », d'entrainer un classifieur ML pour déduire la probabilité de fusion de deux superpixels. Nous avons ainsi, développé une stratégie d'intégration des connaissances dans une approche de segmentation à travers l'apprentissage automatique par RF qui présente deux contributions principales :

1. un mécanisme basé sur l'apprentissage artificiel pour la fusion de superpixels et
2. une technique d'agrégation basée sur des graphes pour réaliser la segmentation d'images.

En particulier, nous avons utilisé le classifieur forêts aléatoires (RF) pour prédire la probabilité de fusion entre toute paire de superpixels voisins. Nous avons choisi la méthode des forêts aléatoires pour apprendre la fusion des superpixels puisqu'elles offrent divers avantages dont l'efficacité de calcul, la robustesse par rapport aux valeurs aberrantes et les résultats probabilistes. Cela a permis de gérer l'écart sémantique dans l'estimation de la similarité et de manipuler des objets avec des parties hétérogènes. Une approche de décomposition multi-niveaux d'images est utilisée pour offrir au classifieur une meilleure capacité d'apprentissage. En outre l'introduction d'une description contextuelle des superpixels dans leur caractérisation permet d'avoir un champ caractéristique plus vaste de ces derniers pour obtenir un apprentissage plus cohérent des similarités.

D'autre part, étant donné que l'image est représentée par une structure de graphe d'adjacence de région pondérée par des probabilités de fusion, la segmentation est réalisée par un algorithme d'agrégation des nuds du graphe. Cet algorithme est basé sur la sélection de la meilleure paire de superpixels fusionnant à chaque itération en utilisant une fonction de similarité objective. Les résultats obtenus sur les jeux de données DAVIS 2017 et ISIC 2018 indiquent une amélioration substantielle de la performance par rapport aux approches de l'état-de-l'art, à la fois visuellement et quantitativement.

Le tableau IV.5 présente un récapitulatif et une mise en perspectives des contributions apportées dans ce travail.

5. Publications scientifiques

L'ensemble des travaux réalisés depuis le début de notre thèse a donné lieu à deux (2) publications scientifiques. La première expose le modèle de la partie intégration des connaissances de l'approche proposée. Elle est publiée en Mars 2016 dans une conférence de classe C. Le travail présenté dans cette

TABLE IV.5. – Tableau des principales contributions par chapitre et leur mise en perspectives.

Contribution	Problématique	Résultats
Segmentation par propagation	Croissance des régions	- Extraction et représentation des connaissances - Focalisation et Croissance par propagation de connaissances possibilistes
Segmentation par croissance adaptative	Regroupement de régions	- Comparaison des régions sur le contenu et la frontière - Ordre de regroupement adaptatif
Similarité par apprentissage automatique	Fossé sémantique	- Similarité par apprentissage RF - Regroupement par sélection de paires de régions - Pondération de similarité par le ratio des régions

conférence portait sur l'utilisation des connaissances a priori pour guider les tâches de la segmentation et d'étiquetage. Nous avons proposé une caractérisation des connaissances par des degrés d'appartenance aux types d'objets dans la scène afin de s'affranchir des descripteurs bas niveaux mais aussi de fournir un cadre de raisonnement flou aux itérations de la phase de segmentation. Une approche de segmentation par croissance des régions guider par les degrés d'appartenance a été proposée et validée par une application sur des mammographies.

La seconde publication propose une stratégie itérative et multi-niveaux de segmentation d'images par croissance des régions en utilisant les superpixels. Elle est publiée dans une revue impactée et disponible en ligne depuis Septembre 2017. Nous avons proposé une extension de l'algorithme de génération de superpixels SLIC (Achanta *et al.*, 2012) pour intégrer les contours. Nous utilisons la technique de croissance des régions sur les superpixels pour la segmentation de l'image. Utiliser des superpixels au lieu des pixels, permet de calculer des meilleurs descripteurs ce qui conduira à une meilleure comparaison des régions et une fusion plus efficace. Notre algorithme utilise des critères sémantiques de fusion de régions dont les valeurs sont mises à jour automatiquement suivant les itérations. Cela permet d'éviter les fusions erronées car les propriétés des régions comparées changent avec les itérations.

1. **Conférence(2016)** : Mahaman Sani Chaibou, Karim Kalti, Basel Solaiman, Mohamed Ali Mahjoub, "A Combined Approach Based on Fuzzy Classification and Contextual Region Growing to Image Segmentation". In *Computer Graphics, Imaging and Visualization (CGiV), 13th International Conference on,* pp. 172-177. IEEE, 2016, **doi : 10.1109/CGiV.2016.41**.
2. **Journal (2017)** : Mahaman Sani Chaibou, Pierre-Henri Conze, Karim Kalti, Basel Solaiman, Mohamed Ali Mahjoub, "Adaptive strategy for superpixel-based region-growing image segmentation," *J. Electron. Imaging* 26(6), 061605 (2017), **doi : 10.1117/1.JEI.26.6.061605**.

6. Perspectives

La problématique traitée dans ce travail de thèse est la segmentation par propagation de connaissances. Elle simule la segmentation de l'expert humain lors de l'analyse d'images. Les travaux présentés dans ce

manuscrit ouvrent plusieurs perspectives.

6.1. Perspectives à court terme

À court terme, nos travaux peuvent avoir des perspectives comme suit.

— Dans le contexte de calcul de similarité entre régions, nous avons proposé une caractérisation multi-échelle des régions. Cette approche peut être étendue pour inclure l'aspect arborescent de la représentation. Cela permettra alors des approches intéressantes pour la comparaison des vecteurs de caractéristiques.

— L'approche de similarité par apprentissage présentent des résultats encourageants. L'adéquation des caractéristiques utilisées pour la description des superpixels mais aussi celle du classifieur d'apprentissage mènera à des prédictions mieux ciblées. En outre, une représentation hiérarchique de la décomposition à plusieurs niveaux en superpixels peut améliorer les performances du classifieur et permettre une intégration aisée et plus robuste du contexte local grâce aux caractéristiques hiérarchiques.

6.2. Perspectives à long terme

Le retour en force de l'intelligence artificielle dans le domaine d'analyse automatique d'images a ramené et entretient un certain espoir dans l'idée des systèmes d'analyse génériques. Même si la tendance actuelle est loin des premiers systèmes, il existe beaucoup de travaux théoriques comme d'implémentation qui s'orientent vers le super-système.

— Nos travaux futurs se focaliseront sur le raisonnement de l'approche proposée dans ce manuscrit. En effet, le travail réalisé couvre seulement les premiers niveaux de représentation de connaissances. L'intégration des niveaux plus hauts ajoutera certainement une précision et une flexibilité dans le raisonnement. Cette expansion va aussi nécessiter des travaux sur la fusion résultats des niveaux ajoutés.

— Un autre aspect fondamental des systèmes à base de connaissances est l'accumulation de l'expérience. En effet, dans la version actuelle de l'approche la base de connaissances ne contient que les connaissances fournies par l'expert au début et celles extraites sur l'image traitée. Une perspective intéressante serait de raffiner les modèles des classes de la base en les mettant à jour à chaque image traitée.

Bibliographie

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua et S. Susstrunk : Slic superpixels. *Dept. School Comput. Commun. Sci., EPFL, Lausanne, Switzerland, Tech. Rep*, 149300, 2010.

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua et S. Süsstrunk : Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34 (11) :2274–2282, 2012.

R. Adams et L. Bischof : Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6) :641–647, 1994.

J. Aloimonos : Purposive and qualitative active vision. *In Pattern Recognition, 1990. Proceedings., 10th International Conference on*, vol. 1, p. 346–360. IEEE, 1990.

J. Aloimonos, I. Weiss et A. Bandyopadhyay : Active vision. *International journal of computer vision*, 1(4) :333–356, 1988.

B. Alsahwa : *Représentation d'un environnement par un système multi-capteurs : fusion et interprétation de scène*. Thèse de doctorat, Télécom Bretagne ; Université de Rennes 1, 2014.

Y. Amit et D. Geman : Shape quantization and recognition with randomized trees. *Neural computation*, 9 (7) :1545–1588, 1997.

N. Audebert, A. Boulch, H. Randrianarivo, B. Le Saux, M. Ferecatu, S. Lefèvre et R. Marlet : Deep learning for urban remote sensing. *In Urban Remote Sensing Event (JURSE), 2017 Joint*, p. 1–4. IEEE, 2017.

M. Baatz et Shäpe : Multi resolution segmentation : an optimum approach for high quality multi scale image segmentation. *In Beutrage zum AGIT-Symposium. Salzburg, Heidelberg, 2000*, p. 12–23, 2000.

R. Bajcsy : Active perception. *Proceedings of the IEEE*, 76(8) :966–1005, 1988.

D. H. Ballard et C. M. Brown : Principles of animate vision. *CVGIP : Image Understanding*, 56(1) :3–21, 1992.

H. G. Barrow et J. M. Tenenbaum : *MSYS : A system for reasoning about scenes*. SRI International Menlo Park CA, 1976.

J. Bergstra et Y. Bengio : Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb) :281–305, 2012.

A. Bhattacharyya : On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35 :99–109, 1943.

A. Bhattacharyya : On a measure of divergence between two multinomial populations. *Sankhyā : the indian journal of statistics*, p. 401–406, 1946.

L. S. Bins, L. G. Fonseca, G. J. Erthal et F. M. Ii : Satellite imagery segmentation : a region growing approach. *Simpósio Brasileiro de Sensoriamento Remoto*, 8(1996) :677–680, 1996.

T. Blaschke et G. J. Hay : Object-oriented image analysis and scale-space : theory and methods for modeling and evaluating multiscale landscape structure. *International Archives of Photogrammetry and Remote Sensing*, 34(4) :22–29, 2001.

R. C. Bolles : Verification vision within a programmable assembly system. Rap. tech., STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1976.

M. Borsotti, P. Campadelli et R. Schettini : Quantitative evaluation of color image segmentation results. *Pattern recognition letters*, 19(8) :741–747, 1998.

A. Bosch, A. Zisserman et X. Munoz : Image classification using random forests and ferns. *In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, p. 1–8. IEEE, 2007.

B. Bouchon-Meunier, M. Rifqi et M.-J. Lesot : Similarities in fuzzy data mining : from a cognitive view to real-world applications. *In IEEE World Congress on Computational Intelligence*, p. 349–367. Springer, 2008.

L. Breiman : Random forests. *Machine learning*, 45(1) :5–32, 2001.

L. Breiman : Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1, 2002.

J. Bückner, M. Pahl, O. Stahlhut et C. Liedtke : Geoaida a knowledge based automatic image data analyser for remote sensing data. *In International ICSC Congress on Computational Intelligence : Methods & Applications*, vol. 2, p. 19–22. Citeseer, 2001.

P. Buyssens, I. Gardin et S. Ruan : Eikonal based region growing for superpixels generation : Application to semi-supervised real time organ segmentation in ct images. *Irbm*, 35(1) :20–26, 2014.

F. Calderero et F. Marques : Region merging techniques using information theory statistical measures. *IEEE transactions on Image Processing*, 19(6) :1567–1586, 2010.

J. Canny : A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6) :679–698, 1986.

G. C. Cawley et N. L. Talbot : Kernel learning at the first level of inference. *Neural Networks*, 53 :69–80, 2014.

M. E. Clarkson : Intelligent user interface for the detection of arbitrary shapes by mathematical morphology. *In Image Algebra and Morphological Image Processing III*, vol. 1769, p. 82–94. International Society for Optics and Photonics, 1992.

D. Comaniciu et P. Meer : Mean shift : A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5) :603–619, 2002.

P.-H. Conze, V. Noblet, F. Rousseau, F. Heitz, V. De Blasi, R. Memeo et P. Pessaux : Scale-adaptive supervoxel-based random forests for liver tumor segmentation in dynamic contrast-enhanced CT scans. *International journal of computer assisted radiology and surgery*, 12(2) :223–233, 2017.

D. Crevier et R. Lepage : Knowledge-based image understanding systems : A survey. *Computer Vision and Image Understanding*, 67(2) :161–185, 1997.

D. J. Crisp, P. Perry et N. J. Redding : Fast segmentation of large images. *In Proceedings of the 26th Australasian computer science conference-Volume 16*, p. 87–93. Australian Computer Society, Inc., 2003.

A. Darwish, K. Leukert et W. Reinhardt : Image segmentation for the purpose of object-based classification. *In Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International*, vol. 3, p. 2039–2041. Ieee, 2003.

R. C. De Amorim et B. Mirkin : Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, 45(3) :1061–1075, 2012.

R. Deriche : Optimal edge detection using recursive filtering. *International Journal of Computer Vision*, 2 :167–187, 1987.

L. R. Dice : Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297–302, 1945.

B. A. Draper, R. T. Collins, J. Brolio, A. R. Hanson et E. M. Riseman : The schema system. *International Journal of Computer Vision*, 2(3) :209–250, 1989.

D. Dreizin, U. K. Bodanapally, N. Neerchal, N. Tirada, M. Patlas et E. Herskovits : Volumetric analysis of pelvic hematomas after blunt trauma using semi-automated seeded region growing segmentation : a method validation study. *Abdominal Radiology*, 41(11) :2203–2208, 2016.

F. Drucker et J. MacCormick : Fast superpixels for video analysis. *In Workshop on Motion and Video Computing*, p. 1–8, 2009.

D. Dubois et H. Prade : Unfair coins and necessity measures : towards a possibilistic interpretation of histograms. *Fuzzy sets and systems*, 10(1-3) :15–20, 1983.

D. Dubois et H. Prade : Possibility theory : An approach to the computerized processing of information, 1988.

D. J. Dubois : *Fuzzy sets and systems : theory and applications*, vol. 144. Academic press, 1980.

L. D. Erman, F. Hayes-Roth, V. R. Lesser et D. R. Reddy : The hearsay-ii speech-understanding system : Integrating knowledge to resolve uncertainty. *ACM Computing Surveys (CSUR)*, 12(2) :213–253, 1980.

A.-M. Forte, M. Bernadet, F. Lavaire et Y. J. Bizais : Object-oriented versus logical conventional implementation of a mmiis. *In Medical Imaging VI*, p. 215–224. International Society for Optics and Photonics, 1992.

J. Freixenet, X. Muñoz, D. Raba, J. Martí et X. Cufí : Yet another survey on image segmentation : Region and boundary information integration. *In European Conference on Computer Vision*, p. 408–422. Springer, 2002.

H. Fu, X. Cao, D. Tang, Y. Han et D. Xu : Regularity preserved superpixels and supervoxels. *IEEE Transactions on Multimedia*, 16(4) :1165–1175, 2014.

G. Gan, C. Ma et J. Wu : *Data clustering : theory, algorithms, and applications*, vol. 20. Siam, 2007.

P. Garnesson, G. Giraudon et P. Montesinos : MESSIE : un systeme multispécialistes en vision. applicationa linterprétation en imagerie aérienne, 1989.

P. Garnesson, G. Giraudon et P. Montesinos : MESSIE : un systeme multispécialistes en vision. application a linterprétation en imagerie aérienne. *Traitement du signal*, 9(5) :403–419, 1992.

J. Gibson : The ecological approach to visual perception houghton mifflin boston google scholar. 1979.

G. Giraudon : Chaînage efficace de contours. Rap. tech. 605, INRIA, 2 1987.

G. Giraudon et P. Montesinos : Coopération contour-région pour la détection infrarouge. *6e Congrès RFIA, Antibes*, 11 1987.

P. M. Granitto, C. Furlanello, F. Biasioli et F. Gasperi : Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2) :83–90, 2006.

L. Guigues : *Modèles multi-échelles pour la segmentation d'images*. Thèse de doctorat, Cergy-Pontoise, 2003.

I. Guyon, J. Weston, S. Barnhill et V. Vapnik : Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3) :389–422, 2002.

M. Hadjieleftheriou et D. Srivastava : Weighted set-based string similarity. *IEEE Data Eng. Bull.*, 33 (1) :25–36, 2010.

A. Hajdu et T. Tóth : Approximating non-metrical minkowski distances in 2D. *Pattern Recognition Letters*, 29(6) :813–821, 2008.

R. Hamming : The bell system technical journal. *Bell Syst. Tech. J*, 26(2) :147–160, 1950.

R. M. Haralick, K. Shanmugam *et al.* : Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6) :610–621, 1973.

R. M. Haralick et L. G. Shapiro : Image segmentation techniques. *Computer vision, graphics, and image processing*, 29(1) :100–132, 1985.

P. Harmon et D. King : *Expert systems*. John Wiley & Sons, Inc., 1985.

B. Hayes-Roth : *The blackboard architecture : A general framework for problem solving ?* Heuristic Programming Project, Computer Science Department, Stanford University, 1983.

B. Hayes-Roth : A blackboard architecture for control. *Artificial intelligence*, 26(3) :251–321, 1985.

C.-Y. Hsu et J.-J. Ding : Efficient image segmentation algorithm using slic superpixels and boundary-focused region merging. *In Information, Communications and Signal Processing (ICICS) 2013 9th International Conference on*, p. 1–5. IEEE, 2013.

Z. Huang, X. Wang, J. Wang, W. Liu et J. Wang : Weakly-supervised semantic segmentation network with deep seeded region growing. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 7014–7023, 2018.

C. Hudelot : *Towards a cognitive vision platform for semantic image interpretation ; application to the recognition of biological organisms*. Thèse de doctorat, Université Nice Sophia Antipolis, 2005.

C. Hudelot et M. Thonnat : A cognitive vision platform for automatic recognition of natural complex objects. *In Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*, p. 398–405. IEEE, 2003.

A. Humayun, F. Li et J. M. Rehg : The middle child problem : Revisiting parametric min-cut and seeds for object proposals. *In Proceedings of the IEEE International Conference on Computer Vision*, p. 1600–1608, 2015.

M. Ichino et H. Yaguchi : Generalized minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(4) :698–708, 1994.

K. Ikeuchi et T. Kanade : Automatic generation of object recognition programs. *Proceedings of the IEEE*, 76(8) :1016–1035, 1988.

H. Iwase, T. Toriu et T. Gotoh : An expert system for image processing. *In Artificial Intelligence Applications, 1988., Proceedings of the Fourth Conference on*, p. 395–399. IEEE, 1988.

P. Jaccard : Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37 :241–272, 1901.

A.-L. Jousselme et P. Maupin : Distances in evidence theory : Comprehensive survey and generalizations. *International Journal of Approximate Reasoning*, 53(2) :118–145, 2012.

T. Kanade : Model representations and control structures in image understanding. *In IJCAI*, p. 1074–1082, 1977.

S. Kullback et R. A. Leibler : On information and sufficiency. *The annals of mathematical statistics*, 22 (1) :79–86, 1951.

P. Lassalle, J. Inglada, J. Michel, M. Grizonnet et J. Malik : A scalable tile-based framework for region-merging segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10) :5473–5485, 2015.

V. Lepetit et P. Fua : Keypoint recognition using randomized trees. *IEEE transactions on pattern analysis and machine intelligence*, 28(9) :1465–1479, 2006.

M. D. Levine *et al.* : A knowledge-based computer vision system. *Computer vision systems*, 78 :335–352, 1978.

A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson et K. Siddiqi : Turbopixels : Fast superpixels using geometric flows. *IEEE transactions on pattern analysis and machine intelligence*, 31(12) :2290–2297, 2009.

Z. Li et J. Chen : Superpixel segmentation using linear spectral clustering. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 1356–1363, 2015.

C. Liedtke, J. Bückner, O. Grau, S. Growe et R. Tönjes : AIDA : A system for the knowledge based interpretation of remote sensing data. *In 3rd International Airborne Remote Sensing Conference*, vol. 7, 1997.

C. Liedtke, J. Bückner, M. Pahl et O. Stahlhut : Knowledge based system for the interpretation of complex scenes. *Automatic Extraction of Man-Made Objects from Aerial and Space Images (III), AA Balkema Publishers, Lisse/Abingdon/Exton (PA)/Tokio*, p. 3–12, 2001.

G. Louppe : Understanding random forests : From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

V. Machairas, T. Baldeweck, T. Walter et E. Decenciere : New general features based on superpixels for image segmentation learning. *In Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, p. 1409–1413. IEEE, 2016.

V. Machairas, M. Faessel, D. Cárdenas-Peña, T. Chabardes, T. Walter et E. Decencière : Waterpixels. *IEEE Transactions on Image Processing*, 24(11) :3707–3716, 2015.

N. Maillot, M. Thonnat et C. Hudelot : Ontology based object learning and recognition : Application to image retrieval. *In Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, p. 620–625. IEEE, 2004.

D. Marr : *Vision : A computational investigation into the human representation and processing of visual information*. W H Freeman and Company, New York, 1982.

D. Martin, C. Fowlkes, D. Tal et J. Malik : A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *In null*, p. 416. IEEE, 2001.

T. Matsuyama et V. S.-S. Hwang : SIGMA : A framework for image understanding-integration of bottom-up and top-down analysis. *In IJCAI*, vol. 9, p. 908–915, 1985.

T. Matsuyama et V. S.-S. Hwang : *SIGMA : A knowledge-based aerial image understanding system*. Springer Science & Business Media, 1990.

D. M. McKeown et J. L. Denlinger : *Map-guided feature extraction from aerial imagery*. Carnegie-Mellon University, Department of Computer Science, 1984.

D. M. McKeown, W. A. Harvey et J. McDermott : Rule-based interpretation of aerial imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5) :570–585, 1985.

A. Mehnert et P. Jackway : An improved seeded region growing algorithm. *Pattern Recognition Letters*, 18(10) :1065–1071, 1997.

M. Meilă : Comparing clusterings an information based distance. *Journal of multivariate analysis*, 98 (5) :873–895, 2007.

J. M. Merigo et M. Casanovas : A new minkowski distance based on induced aggregation operators. *International Journal of Computational Intelligence Systems*, 4(2) :123–133, 2011.

M. Mignotte : A label field fusion model with a variation of information estimator for image segmentation. *Information Fusion*, 20 :7–20, 2014.

M. Minsky : A framework for representing knowledge. 1975.

P. Neubert et P. Protzel : Superpixel benchmark and comparison. *In Proc. Forum Bildverarbeitung*, p. 1–12, 2012.

P. Neubert et P. Protzel : Compact watershed and preemptive slic : On improving trade-offs of superpixel segmentation algorithms. *In Pattern Recognition (ICPR), 2014 22nd International Conference on*, p. 996–1001. IEEE, 2014.

A. Newell : Production systems : Models of control structures. Rap. tech., CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1973.

H. Niemann, G. F. Sagerer, S. Schroder et F. Kummert : Ernest : A semantic network system for pattern understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9) :883–905, 1990.

S. Nikolopoulos, G. T. Papadopoulos, I. Kompatsiaris et I. Patras : Evidence-driven image interpretation by combining implicit and explicit knowledge in a bayesian network. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(5) :1366–1381, 2011.

D. Oneata, J. Revaud, J. Verbeek et C. Schmid : Spatio-temporal object detection proposals. *In European conference on computer vision*, p. 737–752. Springer, 2014.

M. Ozuysal, P. Fua et V. Lepetit : Fast keypoint recognition in ten lines of code. *In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, p. 1–8. Ieee, 2007.

E. Parzen : On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3) :1065–1076, 1962.

B. S. Paskaleva et P. B. Bochev : A vector space model for information retrieval with generalized similarity measures. Rap. tech., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2012.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot et E. Duchesnay : Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

B. Peng, L. Zhang et D. Zhang : Automatic image segmentation by dynamic region merging. *IEEE Transactions on image processing*, 20(12) :3592–3605, 2011.

J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung et L. Van Gool : The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

A. Popescu, P.-A. Moëllic et C. Millet : Semretriev : an ontology driven image retrieval system. *In Proceedings of the 6th ACM international conference on Image and video retrieval*, p. 113–116. ACM, 2007.

A. R. Rao et R. Jain : Knowledge representation and control in computer vision systems. *IEEE Expert : Intelligent Systems and Their Applications*, 3(1) :64–79, 1988.

R. Reiter et A. K. Mackworth : A logical framework for depiction and image interpretation. *Artificial Intelligence*, 41(2) :125–155, 1989.

X. Ren et J. Malik : Learning a classification model for segmentation. *In null*, p. 10. IEEE, 2003.

C. Rosenberger et K. Chehdi : Genetic fusion : application to multi-components image segmentation. *In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 4, p. 2223–2226. IEEE, 2000.

F. Sandakly et G. Giraudon : Scene analysis system. *In Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, vol. 3, p. 806–810. IEEE, 1994.

N. Sangsefidi, A. H. Foruzan et A. Dolati : Balancing the data term of graph-cuts algorithm to improve segmentation of hepatic vascular structures. *Computers in biology and medicine*, 2017.

T. M. Santana, A. M. Machado, A. d. A. Araújo et J. A. dos Santos : Star : A contextual description of superpixels for remote sensing image classification. *In Iberoamerican Congress on Pattern Recognition*, p. 300–308. Springer, 2016.

R. Schettini : A segmentation algorithm for color images. *Pattern Recognition Letters*, 14(6) :499–506, 1993.

A. Sharon, G. Meirav, B. Ronen et B. Achi : "image segmentation by probabilistic bottom-up aggregation and cue integration.". *In "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition"*, June 2007.

J. Shi et J. Malik : Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8) :888–905, 2000.

F. Y. Shih et S. Cheng : Automatic seeded region growing for color image segmentation. *Image and vision computing*, 23(10) :877–886, 2005.

R. E. Silva : An alternative approach to counting minimum (s ; t)-cuts in planar graphs. 2017.

A. W. Smeulders, M. Worring, S. Santini, A. Gupta et R. Jain : Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12) :1349–1380, 2000.

G. W. Snedecor et W. G. Cochran : Statistical methods. ames, 1967.

T. Sørensen : A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5 :1–34, 1948.

D. Stutz, A. Hermans et B. Leibe : Superpixels : an evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, 166 :1–27, 2018.

J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok *et al.* : The mammographic image analysis society digital mammogram database. *In Exerpta Medica. International Congress Series*, vol. 1069, p. 375–378, 1994.

T. Thirumeni, R. John et S. Shaikh : 3d segmentation of glioma from brain mr images using seeded region growing and fuzzy c-means clustering. *image*, 170 :255, 2015.

M. Tkalcic et J. F. Tasic : *Colour spaces : perceptual, historical and applicational background*, vol. 1. IEEE, 2003.

C. Town : Ontological inference for image and video analysis. *Machine Vision and Applications*, 17 (2) :94–115, 2006.

A. Tremeau et N. Borel : A region growing and merging algorithm to color segmentation. *Pattern recognition*, 30(7) :1191–1203, 1997.

Z. Tu et S.-C. Zhu : Image segmentation by data-driven markov chain monte carlo. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5) :657–673, 2002.

A. Tversky : Features of similarity. *Psychological review*, 84(4) :327, 1977.

R. Unnikrishnan, C. Pantofaru et M. Hebert : Toward objective evaluation of image segmentation algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 29(6), 2007.

M. Van den Bergh, X. Boix, G. Roig et L. Van Gool : Seeds : Superpixels extracted via energy-driven sampling. *International Journal of Computer Vision*, 111(3) :298–314, 2015.

J. E. Vargas, A. X. Falcão, J. Dos Santos, J. C. D. M. Esquerdo, A. C. Coutinho et J. Antunes : Contextual superpixel description for remote sensing image classification. *In 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, p. 1132–1135. IEEE, 2015.

A. Vedaldi et S. Soatto : Quick shift and kernel methods for mode seeking. *In European Conference on Computer Vision*, p. 705–718. Springer, 2008.

H. Vincent, Shang-Shouq, D. Larry, S. et M. Takashi : Hypothesis integration in image understanding systems. *Traitement du signal*, 9(5) :239–305, 1985.

H. Wache, T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann et S. Hübner : Ontology-based integration of information-a survey of existing approaches. *In IJCAI-01 workshop : ontologies and information sharing*, vol. 2001, p. 108–117. Seattle, USA, 2001.

M. Wang, X. Liu, Y. Gao, X. Ma et N. Q. Soomro : Superpixel segmentation : a benchmark. *Signal Processing : Image Communication*, 56 :28–39, 2017.

J. H. Ward Jr : Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244, 1963.

P. H. Winston : Learning structural descriptions from examples. 1970.

A. Y. Yang, J. Wright, Y. Ma et S. S. Sastry : Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 110(2) :212–225, 2008.

Y. Yang, Y. Wang et X. Xue : A novel spectral clustering method with superpixels for image segmentation. *Optik-International Journal for Light and Electron Optics*, 127(1) :161–167, 2016.

J. Yao, M. Boben, S. Fidler et R. Urtasun : Real-time coarse-to-fine topologically preserving segmentation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 2947–2955, 2015.

P. Yin, A. Criminisi, J. Winn et I. Essa : Tree-based classifiers for bilayer video segmentation. *In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, p. 1–8. IEEE, 2007.

S. Yin, Y. Qian et M. Gong : Unsupervised hierarchical image segmentation through fuzzy entropy maximization. *Pattern Recognition*, 2017.

H. Yu, X. Zhang, S. Wang et B. Hou : Context-based hierarchical unequal merging for sar image segmentation. *IEEE Transactions on geoscience and remote sensing*, 51(2) :995–1009, 2013.

Q. Yu et D. A. Clausi : IRGS : Image segmentation using edge penalties and region growing. *IEEE transactions on pattern analysis and machine intelligence*, 30(12) :2126–2139, 2008.

L. A. Zadeh : Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 1(1) :3–28, 1978.

H. Zhang, J. E. Fritts et S. A. Goldman : An entropy-based objective evaluation method for image segmentation. *In Storage and Retrieval Methods and Applications for Multimedia 2004*, vol. 5307, p. 38–50. International Society for Optics and Photonics, 2003.

H. Zhang, J. E. Fritts et S. A. Goldman : Image segmentation evaluation : A survey of unsupervised methods. *computer vision and image understanding*, 110(2) :260–280, 2008.

Y. Zhang et K. He : Multi-scale gaussian segmentation via graph cuts. *DEStech Transactions on Computer Science and Engineering*, (csae), 2017.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Segmentation d'image par intégration itérative de connaissances

Mots clés : Image segmentation, Superpixels, Mesure de similarité, Descripteurs contextuels, Descripteurs multi-niveaux, Forêts aléatoires

Résumé : Le traitement d'images est un axe de recherche très actif depuis des années. L'interprétation des images constitue une de ses branches les plus importantes de par ses applications socio-économiques et scientifiques. Cependant cette interprétation, comme la plupart des processus de traitements d'images, nécessite une phase de segmentation pour délimiter les régions à analyser. En fait l'interprétation est un traitement qui permet de donner un sens aux régions détectées par la phase de segmentation. Ainsi, la phase d'interprétation ne pourra analyser que les régions détectées lors de la segmentation.

Bien que l'objectif de l'interprétation automatique soit d'avoir le même résultat qu'une interprétation humaine, la logique des techniques classiques de ce domaine ne marie pas celle de l'interprétation humaine. La majorité des approches classiques d'interprétation d'images séparent la phase de segmentation et celle de l'interprétation. Les images sont d'abord segmentées puis les régions détectées sont interprétées. En plus, au niveau de la segmentation les techniques classiques parcourent les images de manière séquentielle, dans l'ordre de stockage des pixels. Ce parcours ne reflète pas nécessairement le parcours de l'expert humain lors de son exploration de l'image. En effet ce dernier commence le plus souvent par balayer l'image à la recherche d'éventuelles zones d'intérêts. Dans le cas échéant, il analyse les zones potentielles sous trois niveaux de vue pour essayer de reconnaitre de quel objet s'agit-il. Premièrement, il analyse la zone en se basant sur ses caractéristiques physiques. Ensuite il considère les zones avoisinantes de celle-ci et enfin il zoome sur toute l'image afin d'avoir une vue complète tout en considérant les informations locales à la zone et celles de ses voisines.

Pendant son exploration, l'expert, en plus des informations directement obtenues sur les caractéristiques physiques de l'image, fait appel à plusieurs sources d'informations qu'il fusionne pour interpréter l'image. Ces sources peuvent inclure les connaissent acquises grâce à son expérience professionnelle, les contraintes existantes entre les objets de ce type d'images, etc.

L'idée de l'approche présentée ici est que simuler l'activité visuelle de l'expert permettrait une meilleure compatibilité entre les résultats de l'interprétation et ceux de l'expert. Ainsi nous retenons de cette analyse trois aspects importants du processus d'interprétation d'image que nous allons modéliser dans l'approche proposée dans ce travail :

1. Le processus de segmentation n'est pas nécessairement séquentiel comme la plus part des techniques de segmentations qu'on rencontre, mais plutôt une suite de décisions pouvant remettre en cause leurs prédécesseurs. L'essentiel étant à la fin d'avoir la meilleure classification des régions. L'interprétation ne doit pas être limitée par la segmentation.
2. Le processus de caractérisation d'une zone d'intérêt n'est pas strictement monotone i.e. que l'expert peut aller d'une vue centrée sur la zone à vue plus large incluant ses voisines pour ensuite retourner vers la vue contenant uniquement la zone et vice-versa.
3. Lors de la décision plusieurs sources d'informations sont sollicitées et fusionnées pour une meilleure certitude.

La modélisation proposée de ces trois niveaux met particulièrement l'accent sur les connaissances utilisées et le raisonnement qui mène à la segmentation des images.

Title : Image segmentation by iterative knowledge integration

Keywords : Image segmentation, Superpixels, Similarity measure, Region-growing, Contextual features, Multi-level features, Artificial learning, Random forests

Abstract : Image processing has been a very active area of research for years. The interpretation of images is one of its most important branches because of its socio-economic and scientific applications. However, the interpretation, like most image processing processes, requires a segmentation phase to delimit the regions to be analyzed. In fact, interpretation is a process that gives meaning to the regions detected by the segmentation phase. Thus, the interpretation phase can only analyze the regions detected during the segmentation.

Although the ultimate objective of automatic interpretation is to produce the same result as a human, the logic of classical techniques in this field does not marry that of human interpretation. Most conventional approaches to this task separate the segmentation phase from the interpretation phase. The images are first segmented and then the detected regions are interpreted. In addition, conventional techniques of segmentation scan images sequentially, in the order of pixels appearance. This way does not necessarily reflect the way of the expert during the image exploration. Indeed, a human usually starts by scanning the image for possible region of interest. When he finds a potential area, he analyzes it under three viewpoints trying to recognize what object it is. First, he analyzes the area based on its physical characteristics. Then he considers the region's surrounding areas and finally he zooms in on the whole image in order to have a wider view while considering the information local to the region and those of its neighbors.

In addition to information directly gathered from the physical characteristics of the image, the expert uses several sources of information that he merges to interpret the image. These sources include knowledge acquired through professional experience, existing constraints between objects from the images, and so on.

The idea of the proposed approach, in this manuscript, is that simulating the visual activity of the expert would allow a better compatibility between the results of the interpretation and those of the expert.

We retain from the analysis of the expert's behavior three important aspects of the image interpretation process that we will model in this work:

1. Unlike what most of the segmentation techniques suggest, the segmentation process is not necessarily sequential, but rather a series of decisions that each one may question the results of its predecessors. The main objective is to produce the best possible regions classification.
2. The process of characterizing an area of interest is not a one-way process i.e. the expert can go from a local view restricted to the region of interest to a wider view of the area, including its neighbors and vice versa.
3. Several information sources are gathered and merged for a better certainty, during the decision of region characterisation.

The proposed model of these three levels places particular emphasis on the knowledge used and the reasoning behind image segmentation